UNIVERSIDADE FEDERAL DO PARANÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS NUMÉRICOS EM ENGENHARIA

MIGUEL DIOGENES MATRAKAS

REDUÇÃO DE DIMENSIONALIDADE E VISUALIZAÇÃO INTERATIVA DE DADOS MULTIDIMENSIONAIS UTILIZANDO PROCESSAMENTO PARALELO EM GPU

CURITIBA

2016

MIGUEL DIOGENES MATRAKAS

REDUÇÃO DE DIMENSIONALIDADE E VISUALIZAÇÃO INTERATIVA DE DADOS MULTIDIMENSIONAIS UTILIZANDO PROCESSAMENTO PARALELO EM GPU

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia na área de concentração de Programação Matemática, dos setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial à obtenção do grau de Doutor.
Orientador: Dr. Sergio Scheer

CURITIBA

2016

M433 Matrakas, Miguel Diogenes
Redução de dimensionalidade e visualização interativa de dados multidimensionais utilizando processamento paralelo em GPU / Miguel Diogenes Matrakas – Curitiba, 2016.
106 f. : il. color.;

Tese – Universidade Federal do Paraná (UFPR). Setor de Tecnologia. Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE), 2016.

Orientador: Dr. Sergio Scheer

Bibliografia: p. 101 – 105

1. Escala multidimensional. 2. Imagem tridimensional. 3. Processamento paralelo (Computadores). I. Universidade Federal do Paraná (UFPR). II. Scheer, Sergio. III. Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE). IV. Redução de dimensionalidade e visualização interativa de dados multidimensionais utilizando processamento paralelo em GPU.

CDD: 006.693

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor Tecnologia
Programa de Pós Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA
Código CAPES: 40001016030P0

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de **MIGUEL DIOGENES MATRAKAS**, intitulada: "**REDUÇÃO DE DIMENSIONALIDADE E VISUALIZAÇÃO INTERATIVA DE DADOS MULTIDIMENSIONAIS UTILIZANDO PROCESSAMENTO PARALELO EM GPU**", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 29 de Agosto de 2016.

Prof SÉRGIO SCHEER
Presidente da Banca Examinadora (UFPR)

Prof KLAUS DE GEUS
Avaliador Interno (UFPR)

Prof PAULO HENRIQUE SIQUEIRA
Avaliador Interno (UFPR)

Prof FLAVIO BORTOLOZZI
Avaliador Externo (PUC/PR)

Prof CINTHIA OBLADEN DE ALMENDRA FREITAS
Avaliador Externo (PUC/PR)

Dedico este trabalho à minha esposa Alessandra,
que tanto me incentivou e apoiou durante toda esta jornada.

AGRADECIMENTOS

Agradeço principalmente à minha família: Alessandra, Giovanna e Diogenes que durante todo o tempo entenderam minha ausência, e que sem o seu carinho inestimável a execução desta pesquisa não seria possível.

Ao Professor Doutor Sérgio Scheer, um agradecimento especial por ter me aceito como orientando e pela sua competente orientação e contribuições durante a execução deste trabalho.

A todos os membros do Centro de Estudos Avançados em Segurança de Barragens (CEASB) da Fundação Parque Tecnológico de Itaipu (FPTI), em especial a Alexandra da Silva, agradeço o apoio e inestimável auxílio nas questões operacionais durante o andamento do trabalho.

Agradeço a todos os professores do Programa de Pós-Graduação em Métodos Numéricos da Universidade Federal do Paraná que ministraram as disciplinas por mim cursadas, em especial ao Professor Dr. Anselmo Chaves Neto, pelos ensinamentos e lições, que em muito contribuíram com o meu crescimento acadêmico e pessoal, e a Professora Dra. Liliana Gramani pela dedicação à nossa turma.

Para meus colegas da turma do DINTER UFPR-UNIOESTE, agradeço pelo companheirismo, apoio e auxílio nos momentos de dificuldade. Em especial ao Dr. Valmei Abreu Junior pelos esclarecimentos e paciência com as minhas dificuldades; à Dra. Fabiana Frata pelas conversas e esclarecimentos; ao Dr. Geraldo Brito, padrinho de nossa turma, que tanto trabalhou e auxilio a todos com as mais diferentes necessidades.

Agradeço também a Bill Kuo, Wei Wang, Cindy Bruyere, Tim Scheitlin, e Don Middleton do *U.S. National Center for Atmospheric Research* (NCAR), e do *U.S. National Science Foundation* (NSF) por fornecerem os dados da simulação do furacão Isabel a partir do Modelo de Pesquisa e Previsão Meteorológica, utilizados durante a execução do trabalho.

"Computers make excellent and efficient servants,
but I have no wish to serve under them."
- Mr. Spock, Star Trek: The Original Series
(The Ultimate Computer)

"Without followers, evil cannot spread."
- Mr. Spock, Star Trek: The Original Series
(And The Children Shall Lead)

"Insufficient facts always invite danger."
- Mr. Spock, Star Trek: The Original Series
(Space Seed)

"Change is the essential process of all existence."
- Mr. Spock, Star Trek: The Original Series
(Let That Be Your Last Battlefield)

RESUMO

O método de apresentação de um conjunto de dados influencia os processos de análise e tomada de decisão acerca de seu conteúdo. Portanto, o processo de visualização deve representar, da melhor forma possível, as relações existentes entre seus elementos. Fenômenos ou processos reais apresentam conjuntos de dados multidimensionais, para os quais seria ideal utilizar representações visuais com o maior número de características possível, o que nem sempre é viável devido a limitações nos dispositivos e pelo fato de que a compreensão de um conjunto com mais de três dimensões não é natural. O problema abordado é a visualização de um grande conjunto de dados, como os resultantes de simulações numéricas ou do sensoriamento de uma estrutura, processo ou mesmo fenômeno natural a partir de um conjunto de diferentes tipos de sensores, utilizando um ambiente computacional de baixo custo.

Considerando estes casos, são necessárias ferramentas que auxiliem na visualização e análise dos dados produzidos, facilitando sua compreensão pelos distintos profissionais envolvidos. A partir destas considerações, esta pesquisa tem por objetivo propor uma abordagem para realizar a visualização e análise interativas de um volume de dados multidimensional, de modo que todo o conjunto de dados esteja representado na imagem resultante. Para isso utilizar processamento paralelo baseado em processadores gráficos para implementar as técnicas de Redução Dimensional (RD): *Multidimensional Scaling* (MDS) e transformação por Coordenadas Estrela, de modo a produzir imagens que representem o conteúdo do volume multidimensional (*n*-dimensional) de dados.

Quatro abordagens para realizar a visualização de dados multidimensionais são descritas e, posteriormente, testadas em um protótipo utilizando *General-Purpose Computation on Graphics Processing Units* (GPGPU). Os resultados de processamento indicam a viabilidade de se realizar a visualização de um volume de dados *n*-dimensional utilizando uma técnica de RD em um computador de baixo custo equipado com uma placa gráfica.

Palavras-chave: Escala multidimensional, Processamento paralelo, Coordenadas Estrela, Redução Dimensional (RD), Imagem tridimensional.

ABSTRACT

The method of presenting a data set influences the analysis and decision-making processes, about its contents. So the visualization process should represent in the best possible way the different relations between its elements. Phenomena or real processes present multidimensional data sets, for which it would be ideal to use visual representations with as many features as possible, which is not always feasible due to limitations in the devices and the fact that the understanding of a range of more than three dimensions is not natural. The problem addressed is the view of a large data set, as a result of numerical simulations or the sensoring of a structure, process or natural phenomenon from a number of different types of sensors, using for this a low cost computing environment.

Considering these cases, tools are needed to assist in the visualization and analysis of the data produced, facilitating their comprehension by the various professionals involved. Based on these considerations, this research aims to propose an approach to perform interactive visualization and analysis of a multidimensional data volume, so that the entire data set is represented in the resulting image. Using for this parallel processing based on graphical processing units to implement: the MDS and the Star Coordinates transformation Dimensional Reduction (DR) techniques to produce images that represent the contents of the *n*-dimensional data volume.

Four approaches to perform multidimensional data visualization are described and subsequently tested in a prototype using GPGPU. The processing results indicate the feasibility of performing the visualization of a n-dimensional data volume using a DR technique in a low cost computer equipped with a video card

Keywords: Dimensional scale, Parallel processing, Star coordinates, Dimensional Reduction (DR), Tridimensional image.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLAS *Basic Linear Algebra Subprograms*

CCA *Curvilinear Component Analysis*

CIE *Commission Internationale de l'Eclairage*

CMY *Cyan, Magenta, Yellow*

CMYK *Cyan, Magenta, Yellow, Black*

CPU *Central Processing Unit*

cuBLAS *CUDA Basic Linear Algebra Subprograms*

CUDA *Compute Unified Device Architecture ®*

DFD Diagrama de Fluxo de Dados

DR Dimensional Reduction

EDA *Exploratory Data Analysis*

EOM *Equalized Orthogonal Mapping*

FT Função de Transferência

GPGPU *General-Purpose Computation on Graphics Processing Units*

GPU *Graphical Processing Unit*

GTM *Generative Topographic Mapping*

HLLE *Hessian Locally Linear Embedding*

HLS *Hue, L, Saturation*

HSB *Hue, Saturation, Brightness*

HSI *Hue, Saturation, Intensity*

HSV *Hue, Saturation, Value*

HVC *Hue, Value, C*

Isomap *Isometric Feature Mapping*

ITC *Information Theoretical Clustering Analysis*

LLE *Locally Linear Embedding*

LTSA *Local Tangent Space Alignment*

MAO Mapa Auto-Organizável

MDS *Multidimensional Scaling*

MIMD *Multiple Instruction Multiple Data*

MISD *Multiple Instruction Single Data*

MLP *Multi Layer Perceptron*

MST *Minimum Spanning Tree*

MVU *Maximum Variance Unfolding*

NCAR *U.S. National Center for Atmospheric Research*

NLVC *Nonlinear variance conserving approach*

NSF *US National Science Foundation*

nr-MDS *Reduced Neighbors Multidimensional Scaling*

pb-MDS *Particle-based Multidimensional Scaling*

PCA *Principal Component Analysis*

PRSOM *Probabilistic Regularized SOM*

RD Redução Dimensional

RGB *Red, Green, Blue*

SIMD *Single Instruction Multiple Data*

SIMT *Single Instruction Multiple Threads*

SISD *Single Instruction Single Data*

SM *Streamming Multiprocessor*

SMACOF *Scaling by Mojorizing a Complicated Function*

SNE *Stochastic Neighbor Embedding*

SOM *Self-Organizing Map*

SPE *Stochastic Proximity Embedding*

UCP Unidade Central de Processamento

ViSOM *Visualization induced SOM*

SUMÁRIO

1 INTRODUÇÃO

A tomada de decisões a partir de um conjunto de dados é influenciada pela maneira ou método como seus valores são apresentados. Para conjuntos com diferentes relações entre as variáveis devem ser utilizadas técnicas de visualização com as características apropriadas e seus dados devem ser preparados, ou pré-processados, de maneira que possam ser representados corretamente pelo processo de visualização.

Os fenômenos reais normalmente apresentam dados multidimensionais, ou seja, possuem um grande conjunto de características distintas. O ideal seria que estas representações visuais contassem com o máximo possível de dimensões, ou características, para representar o conjunto original de dados, o que nem sempre é possível, pois os dispositivos e sistemas utilizados pelos analistas normalmente representam apenas duas ou três dimensões simultaneamente. Além do que, a compreensão de um conjunto com mais de três dimensões é bastante difícil (WRIGHT, 2007; van der MAATEN; POSTMA; van den HERIK, 2009).

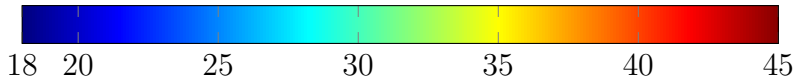
É desejável um sistema ou método mais amigável de apresentar os dados, de modo que seja possível, aos analistas, investigar novas visões do conjunto, tornando mais evidentes as inter-relações existentes. Isto porque na maioria dos casos tem-se gráficos apresentando as variações das grandezas físicas, como apresentado nos trabalhos de Buzzi (2007), Matos (2002) e Patias (2010), o que, em alguns casos, dificulta a observação de relações ou mesmo de divergências entre os valores. Outra dificuldade corresponde à analise da relação existente entre os dados coletados e os valores de referência, provenientes de modelos teóricos, ou ainda em casos de projetos estruturais, ao se comparar os valores simulados, ou resultantes do sensoriamento da estrutura, com os limites de segurança estipulados no momento do projeto da estrutura.

A visualização de um grande conjunto de dados, como é o caso de conjuntos resultantes de simulações numéricas ou seu equivalente resultante do sensoriamento de uma estrutura, processo ou mesmo fenômeno natural a partir de um grande número de sensores, tanto em quantidade como em variedade, é o problema em estudo neste trabalho. A visualização destes conjuntos de dados multidimensionais faz parte da análise de condições estruturais em grandes obras, quer seja durante o seu projeto, execução ou após a finalização da obra, a partir do acompanhamento de sua integridade estrutural. Também são conjunturas relevantes a análise do comportamento em escoamentos de fluidos e o desempenho de um sistema mecânico submetido a diferentes esforços. Pode-se citar também o acompanhamento ou previsão do comportamento de fenômenos naturais como furacões, tornados, entre outros.

Quando se trata de dados com diferentes naturezas físicas, mas que se originam em um mesmo local ou estrutura, as técnicas de visualização não permitem a sua projeção conjunta, ou seja, normalmente em uma mesma imagem que represente o volume em questão não são apresentadas as influências que mais de uma grandeza produzem na estrutura em análise. São utilizadas técnicas que representam cada uma das informações de uma maneira diferente, como uma variação de cores sobreposta por um conjunto de linhas em uma imagem que pode ser tanto bi quanto tridimensional. A visualização dos valores das duas ou mais grandezas são representadas simultaneamente na mesma localização da imagem, ou deslocadas, criando assim um perfil que deve ser interpretado pelo analista.

Para a representação das grandezas de um conjunto de dados existem diferentes métodos e representações possíveis, sendo cada caso particular melhor atendido por uma determinada representação visual. Como exemplo pode-se citar a visualização de uma função que representa a temperatura de um objeto utilizando cores que representam os valores da sua temperatura. Assim uma graduação de cores é definida a partir dos limites do domínio da função, como no exemplo da Figura 1 que apresenta uma escala na qual os valores inicial e final são 18 e 45, respectivamente.

FIGURA 1 – Representação de uma escala de cores



FONTE: Autoria própria

Em uma outra abordagem, o relacionamento entre os valores de diferentes características de um mesmo fenômeno ou estrutura pode fornecer novos indicadores do seu comportamento ao longo de sua evolução. Neste sentido, podem ser citadas algumas das grandezas mais comumente monitoradas em diferentes ramos da engenharia: deslocamentos, tensões, forças, deformações, temperaturas, pressões, vazões e as suas correspondentes direções.

A partir desta lista de grandezas percebe-se que o domínio do problema é multidimensional e portanto as técnicas de visualização a serem aplicadas devem lidar com esta característica para gerar as imagens corretamente, ou seja, evidenciando as áreas ou dados de interesse ao analista do problema em análise, levando-se em consideração as peculiaridades do domínio. Wright (2007) afirma que os conjuntos de dados com muitas variáveis, ou dimensões, são tratados por técnicas de Visualização de Informação e não pela Visualização Científica e em seu artigo Pao e Meng (1998) afirmam que existem evidências que indicam que a redução de dimensões pode ser muito útil para auxiliar os humanos na compreensão de grandes volumes de dados.

Tanto na Visualização Científica quanto na Redução Dimensional podem ser

aplicados algoritmos paralelos (ENGEL et al., 2006; LAWRENCE et al., 2011; PARK; SHIN; HWANG, 2012), ou seja, que aproveitam as características do problema para realizar mais de uma tarefa simultaneamente em sistemas computacionais como as *Graphical Processing Units* (GPUs).

1.1 IMPORTÂNCIA DO TRABALHO

O acompanhamento dos valores mensurados pela instrumentação existente em uma grande obra é o processo de avaliação das condições de sua segurança, sendo realizada por análises estatísticas dos valores mensurados e das relações entre as diferentes variáveis acompanhadas no processo. A aquisição, o processamento e a apresentação dos dados monitorados devem ser feitos levando-se em conta as condições do projeto e os limites estipulados para as medidas de cada grandeza, o que possibilita a identificação de anomalias nestes dados (MATOS, 2002).

Alguns dos principais motivos para a instrumentação e consequente monitoramento de grandes obras são a verificação do projeto, diagnóstico da natureza de eventos adversos e previsão de comportamento da estrutura (MATOS, 2002).

As grandezas monitoradas pela instrumentação das estruturas, no caso de uma barragem de concreto, sofrem influências principalmente dos seguintes fatores: carga direta; subpressões na fundação; pressão intersticial do concreto; calor de hidratação do cimento e sismos. Assim, o monitoramento de deslocamentos, deformações, tensões, temperaturas, níveis piezométricos, pressões e vazões compõem o conjunto de grandezas verificadas, com as análises estatísticas resultando em gráficos que evidenciam o comportamento destas grandezas e/ou os seus relacionamentos (ITAIPU BINACIONAL, 2009; MATOS, 2002; BUZZI, 2007).

Além do acompanhamento de obras e da estabilidade de grandes estruturas, existem também o estudo e análise de eventos naturais, como fenômenos climáticos e sismos, para os quais existem sistemas de coleta de dados em tempo real e a partir destes dados é possível modelar e realizar a simulação matemática que representam estes mesmos eventos. No caso de fenômenos climáticos, o conjunto de grandezas compreende, entre outras, medidas para a quantidade de chuva, neve, granizo, vapor de água, velocidade e direções do vento, e a pressão atmosférica (NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR), 2009).

Considerando os casos das estruturas, dos processos e dos eventos naturais, tanto o seu sensoriamento quanto as suas simulações produzem grandes quantidades de dados para os quais são necessárias ferramentas que auxiliem na sua visualização e análise, viabilizando sua compreensão pelos distintos profissionais envolvidos, uma vez que estes dados são compostos por muitas dimensões (grandezas/variáveis) diferentes e o número de amostras nestes casos normalmente é muito extenso.

Estas características suscitam a questão da possibilidade de se utilizar um computador pessoal para a manipulação e consequente visualização de um volume de dados n-dimensional, indagação esta que, juntamente com as características expostas, fomentam e justificam a proposta deste projeto.

1.2 OBJETIVOS

De acordo com a discussão apresentada, percebe-se a necessidade de elaboração de um método que consiga apresentar de forma satisfatória a projeção de um conjunto de dados multidimensionais em uma visualização volumétrica, para que seja possível a análise das relações e distribuições de valores por parte dos analistas e demais profissionais envolvidos.

Assim a proposta deste trabalho é apresentar um método para, a partir de um conjunto de dados n-dimensional, proveniente de simulações e/ou sensoriamento de uma estrutura, processo ou evento natural, gerar uma imagem representativa contendo simultaneamente todas as diferentes dimensões que o compõem.

1.2.1 Objetivo Geral

O objetivo deste trabalho é desenvolver uma abordagem que permita a redução de dimensões e preparação de dados multidimensionais para a sua visualização e análise interativas utilizando processamento paralelo baseado em processadores gráficos.

De uma forma mais explicativa, nesta pesquisa busca-se desenvolver uma abordagem que permita utilizar programação paralela para GPU, de modo a empregar técnicas de redução de dimensões, para preparar um volume de dados de maneira a permitir que o conjunto seja fornecido como entrada a um algoritmo de visualização de volumes, a partir do qual uma imagem deve ser gerada representando todo o conjunto original de dados, e se possível representando separadamente as classes ou grupos de dados distintos de modo a permitir aos analistas a análise visual e interativa dos relacionamentos entre as classes existentes no volume de dados original.

1.2.2 Objetivos Específicos

Para possibilitar a visualização dos dados conforme a proposta do trabalho são listados a seguir os principais objetivos específicos a serem atingidos:

1. ajustar o volume de dados original, para que os campos escalares de valores possuam as mesmas características de grade, ajustando os espaços, quando necessário, utilizando interpolações;

2. gerar a representação de um conjunto de dados multidimensionais em um espaço que possa ser utilizado como conjunto de entrada em um algoritmo de visualização de volumes - Redução de dimensões;

3. gerar uma representação visual do conjunto de dados a partir de sua transformação, ou projeção, em um espaço de dimensões reduzidas (Visualização), permitindo que parâmetros do processo de obtenção da representação visual do conjunto de dados sejam alterados de forma interativa, de maneira a ajustar a imagem às necessidades dos analistas;

4. criar uma codificação, que seja compatível com a redução de dimensões e com a entrada do algoritmo de visualização, para representar as diferentes classes, ou agrupamentos, existentes nos dados originais, de maneira que estas classes sejam percebidas na imagem resultado - Codificação;

5. propor uma solução para a execução dos algoritmos, aplicando processamento paralelo em equipamentos de baixo custo ou de uso pessoal.

Considerando-se os objetivos estabelecidos, espera-se, ao final deste trabalho, a definição de uma técnica para a visualização de volumes de dados, que possa ser utilizada em um ambiente computacional de baixo custo. Para isso, sendo necessária a implementação de um protótipo para a realização de testes, de modo a confirmar ou não a factibilidade de se realizar a visualização dos dados conforme especificado.

Também deverá estar disponível um conjunto de volumes de dados, a serem utilizados nos testes, com diferentes características. Estes dados podem ser provenientes de distintas fontes, como resultados de simulações numéricas, coletas a partir de sensores, ou mesmo artificialmente compostos, com o objetivo de testar diferentes características dos algoritmos a serem desenvolvidos, bem como das suas soluções.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está organizado em capítulos que aprofundam os conhecimentos necessários ao entendimento completo do problema exposto nas seções anteriores. No Capítulo 2 será apresentada uma revisão dos trabalhos mais relevantes que abordam ou utilizam os métodos e algoritmos considerados na solução do problema em questão, ou por outro lado, foram fonte de motivação para o desenvolvimento da presente pesquisa. Os principais tópicos de interesse nesta revisão são: a visualização de volumes, a visualização de dados multidimensionais, a análise de dados multidimensionais, a redução do número de dimensões de um conjunto de dados, e a computação genérica utilizando processadores gráficos.

Os elementos utilizados durante o desenvolvimento do trabalho são elencados e descritos no Capítulo 3, explicitando os equipamentos, sistemas, métodos ou algoritmos, bem como as bases de dados consideradas para a execução dos testes dos algoritmos desenvolvidos durante a realização desta pesquisa.

Na sequência, o Capítulo 4 apresenta os elementos e algoritmos utilizados e desenvolvidos para esta pesquisa com o intuito de resolver a proposta apresentada na Seção 1.2. Os itens abordados descrevem não somente a estrutura lógica dos algoritmos, mas também as suas estruturas de dados tanto de entrada como de saída. As seções separam o conteúdo em projeções unidimensionais e bidimensionais, fazendo uso da redução dimensional por escala multidimensional ou por coordenadas estrela.

A descrição das saídas obtidas com a execução destes algoritmos está explicitada no Capítulo 5, juntamente com uma análise dos resultados alcançados, evidenciando pontos promissores e elementos que necessitam mais estudos e melhorias para apresentarem um resultado satisfatório.

Por último, no Capítulo 6 estão dispostas as considerações finais a respeito do trabalho, bem como, a partir dos elementos discutidos no Capítulo 5, os pontos promissores e itens que necessitam maiores estudos e melhorias para apresentarem saídas satisfatórias, gerando sugestões de trabalhos futuros para aprimorar os resultados obtidos.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentadas, de forma sucinta, as técnicas relevantes ao desenvolvimento do trabalho proposto, conforme as descrições encontradas em trabalhos correlatos, não pretendendo exaurir os temas, mas sim colocar de forma sucinta as bases necessárias para o entendimento da presente proposta.

Por se tratar de um assunto multidisciplinar, os conteúdos estão separados em seções, iniciando-se pela visualização de volumes, na qual é apresentada a teoria e fundamentação matemática a respeito da geração de imagens a partir de um volume de dados numéricos. Em seguida as seções de análise e visualização de dados multidimensionais abordam os métodos empregados para a compreensão e apresentação do conteúdo de conjuntos cujos elementos possuem um grande número de dimensões. A seção seguinte trata da redução dimensional, abordando os aspectos teóricos e diversos algoritmos utilizados para tratar conjuntos de dados de maneira a possibilitar tanto a compreensão quanto a visualização destes em uma série de aplicações distintas. Na última seção é apresentada uma descrição das características, tanto físicas quanto de utilização, de processadores gráficos fornecendo um referencial para o ambiente computacional a ser utilizado no desenvolvimento deste trabalho.

2.1 VISUALIZAÇÃO

A área de visualização teve sua origem nos anos 1980, a partir de um relatório da *US National Science Foundation* (NSF), quando os computadores passaram a ter capacidade de processar grandes quantidades de dados e exibir os resultados de forma gráfica nos terminais dos cientistas (WRIGHT, 2007). As mesmas técnicas são utilizadas para gerar imagens para a visualização de dados científicos e para a produção de efeitos especiais, ou criação de imagens realistas contendo nuvens, fumaça, fogo ou outros elementos com as mesmas características utilizados na industria do entretenimento, como em filmes ou jogos eletrônicos, isto porque os dois campos baseiam-se nos mesmos modelos físicos, utilizando os mesmos algoritmos para a geração das representações gráficas desejadas (ENGEL et al., 2006).

Wright (2007) definem Visualização como um processo interativo para entender o que gerou, ou produziu, os dados, e não apenas uma técnica de apresentação destes dados. Afirma também que o ser humano compreende naturalmente três dimensões, consequentemente, a compreensão de espaços com maior número de dimensões, com exceção do caso especial do tempo, é limitada. Portanto, se for necessário representar mais variáveis do que podem ser acomodadas com estas restrições, outros recursos devem ser utilizados, como cores, sons, animação, ou o que mais estiver

disponível.

Já Ward, Grinstein e Keim (2015) definem visualização como a comunicação de informação usando representações gráficas, informando também que as figuras são utilizadas para comunicação antes mesmo da formalização da linguagem escrita, e que o processamento da informações gráficas é muito mais rápido que a compreensão de um texto, levando em consideração que as imagens não dependem da língua adotada.

A respeito de Visualização Científica de Dados e Visualização de Informações, Ward, Grinstein e Keim (2015) destacam que, a princípio, a comunidade científica faz distinção entre ambos, mas em seu trabalho esta diferenciação não é aplicada, pois em ambos os casos são produzidas representações visuais dos dados, porém os conjuntos de origem possuem diferentes características.

Nos próximos tópicos são abordadas as técnicas e algoritmos utilizados para realizar a visualização de grandes volumes de dados.

2.1.1 Renderização de volumes

Segundo Engel et al. (2006) a geração de representações gráficas de volumes necessita que o meio participante seja modelado juntamente com o mecanismo de transporte de energia luminosa, e tanto a representação de fenômenos gasosos, quanto a visualização científica de dados volumétricos compartilham o mesmo mecanismo de propagação da energia luminosa. Neste trabalho, estas definições são concernentes ao que se refere à visualização de um volume de dados escalares.

No modelo utilizado para realizar a renderização de volumes, assume-se que a luz se propaga em linhas retas caso não haja interação com o meio. Conforme apresentado nos trabalhos de Engel et al. (2006), Glassner (1995), os três principais tipos de interação que podem ocorrer entre um raio de luz e o meio pelo qual ele está se propagando são:

Emissão: caso no qual o material efetivamente emite luz, aumentando a quantidade de energia que se propaga no meio. Este é o caso, por exemplo, de gases aquecidos, que convertem calor em radiação ou energia luminosa.

Absorção: ocorre quando o material pelo qual o raio de luz está viajando consegue converter energia radiativa em calor, efetivamente diminuindo a quantidade de energia luminosa.

Dispersão: é a situação na qual a direção do raio luminoso é alterada pelo meio que ele está atravessando. Esta alteração pode ser *elástica*, caso no qual o comprimento de onda da radiação luminosa não é alterado pela mudança de

direção, ou *inelástica*, quando existe uma alteração no comprimento de onda da luz. A dispersão também pode ser aditiva, quando a direção de outro raio luminoso é alterada e coincide com a direção do raio atual, aumentando sua energia, ou subtrativa, quando parte da energia do raio luminoso é desviada para outra direção.

Segundo Engel et al. (2006) a energia de um raio de luz pode ser descrita por sua radiância I, que é definida pela quantidade de energia radiativa Q por unidade de área A, que é medida na projeção ao longo da direção de propagação do raio luminoso indicado por $\perp$, pelo ângulo do sólido Ω e por unidade de tempo t:

$$I = \frac{dQ}{dA_{\perp}\, d\Omega\, dt}. \tag{1}$$

Ao se considerar os efeitos da emissão, absorção e dispersão, tem-se que a quantidade de energia luminosa é aumentada pelos efeitos da emissão e da dispersão aditiva, ou é diminuída pelos efeitos da absorção e da dispersão subtrativa.

O modelo tradicional de visualização de dados provenientes de simulações consiste na geração de uma matriz com os dados e, em seguida, realizar a visualização usando métodos tradicionais baseados em interpolação linear. Os autores Nelson, Kirby e Haimes (2014) propuseram um método para realizar a visualização deste tipo de dados de forma precisa e interativa.

O problema reside em calcular ou obter uma aproximação do valor da integral de renderização de volume, que não possui solução analítica e portanto precisa ser resolvida por técnicas numéricas, que introduzem erros na imagem resultante ou levam muito tempo para serem calculadas.

O modelo ótico do método utilizado por Nelson, Kirby e Haimes (2014) é o emissão-absorção (*emission-absorption*), para o qual a irradiação ao longo de um segmento de raio é dada por:

$$I(a,b) = \int_a^b k\left(f(t)\right) \tau\left(f(t)\right) e^{-\int_a^t \tau(f(u))\, du}\, dt, \tag{2}$$

onde a e b são os limites do segmento e k e τ são a cor e a função de transferência de densidade, $f(t)$ é o campo escalar no ponto de descontinuidade t ao longo do segmento que representa o raio de luz. Dada a função f, a sua composição com a função de transferência (convolução) resulta em uma função contínua e derivável apenas em um conjunto finito de pontos de descontinuidade. O cálculo da integral de visualização requer o conhecimento destes pontos, tomando a seguinte forma, com t_i sendo o i-ésimo ponto de descontinuidade:

$$I = \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} k\left(f(t)\right) \tau\left(f(t)\right) e^{-\sum_{j=0}^{i-1} \int_{t_j}^{t_{j+1}} \tau(f(u))\, du - \int_{t_i}^{t} \tau(f(u))\, du}\, dt, \tag{3}$$

que, apesar de ser possível e atraente, na prática é ineficiente devido ao fato que a determinação dos *n* pontos de descontinuidade é equivalente ao cálculo para encontrar as isosuperfícies associadas a cada descontinuidade. A convergência de métodos de quadratura de alta ordem assume funções suaves (*smooth* - funções que possuem derivadas de todas as ordens), o que é violado pelos pontos de descontinuidade.

No método proposto por Nelson, Kirby e Haimes (2014), as funções são classificadas quanto à estrutura do campo sobre o segmento do raio, sendo considerados: segmentos vazios, quando a função de transferência é zero em toda a extensão do segmento; segmentos suaves por partes (*piecewise-smooth*), quando o segmento é suave por partes, possuindo um ou mais pontos de descontinuidade; e segmentos suaves, quando o segmento não possui pontos de descontinuidade.

Para realizar a classificação dos segmentos é utilizada aritmética de intervalos, o que reduz o tempo de processamento.

O cálculo da quadratura de $f(t)$ em $[a, b]$ com um ponto de descontinuidade $c \in [a, b]$,

$$f(t) = \begin{cases} e(t), & t \leq c, \\ g(t), & t > c, \end{cases} \tag{4}$$

com $e(t) = g(t)$, $e'(t) \neq g'(t)$, $e \in C^{\infty}[a, b]$ e $g \in C^{\infty}[a, b]$. Considerando métodos de quadratura da forma:

$$\int_a^b f(t)\, dt \approx \sum_{i=1}^{n} w_i f(t_i) \tag{5}$$

onde $a \leq t_i \leq b$, e $\sum_{i=1}^{n} w_i = b - a$.

Engel et al. (2006) apresentaram o modelo físico do transporte da luz na forma da Equação (6), na qual estão explicitados o coeficiente de absorção do meio, o coeficiente de emissão, e os coeficientes de dispersão tanto aditivo quanto subtrativo.

$$\begin{aligned} w \cdot \nabla_x I(x, w) = \; & -\left(k(x, w) + \sigma(x, w)\right) I(x, w) + q(x, w) \\ & + \int_{esfera} \sigma(x, w')p(x, w', w)I(x, w')d\Omega' \end{aligned} \tag{6}$$

Para lidar com a questão correspondente ao volume de cálculo necessário para resolver o modelo de transporte da luz apresentado, Engel et al. (2006) apresentaram estratégias onde um ou mais termos da equação são removidos ou simplificados:

Apenas absorção: o volume é constituído de material frio e negro, de tal maneira que nenhuma luz é refletida ou dispersa.

Apenas emissão: o volume é constituído de gás que apenas emite luz e é completamente transparente, e portanto a absorção e a dispersão são ignoradas.

Modelo de emissão e absorção: modelo mais comum na visualização de volumes, no qual o material emite e absorve energia luminosa, mas a dispersão e a iluminação indireta são ignoradas.

Dispersão simples e sombreamento: a luz incidente no volume sofre dispersão, e a luz proveniente de fontes externas é atenuada, formando assim áreas de sombra.

Dispersão múltipla: neste caso são levados em consideração todos os termos da Equação (6).

Engel et al. (2006) definiram ainda a equação para visualização de volumes como:

$$I(D) = I_0 e^{-\int_{s_0}^{D} k(t)dt} + \int_{s_0}^{D} q(s) e^{-\int_{s}^{D} k(t)dt} ds \ , \tag{7}$$

na qual I_0 representa a luz que entra no volume pelo ponto $s = s_0$, $I(D)$ representa a quantidade de luz que sai do volume pelo ponto $s = D$ e chega até a câmera. O conjunto de operações para gerar uma imagem representativa de um volume de dados é composto por:

Travessia dos dados: definição ou escolha dos pontos no volume de dados, serve de base para a discretização da integral de visualização do volume contínuo.

Interpolação: normalmente os pontos de amostragem são diferentes da grade, portanto, é necessário reconstruir o espaço contínuo a partir da grade para obter-se os valores das amostras.

Cálculo do gradiente: o gradiente de um campo escalar normalmente é utilizado na determinação da iluminação local.

Classificação: realizada normalmente por funções de transferência, é utilizada para mapear propriedades dos dados em características óticas, normalmente como um conjunto de valores de cor e opacidade.

Iluminação e sombreamento: o sombreamento do volume pode ser incorporado ao processo pela adição de um termo de iluminação na integral de visualização, representada na Equação (7).

Composição: é o processo iterativo para determinar o valor da integral de visualização, que pode realizar o cálculo tanto partindo do observador quanto chegando neste.

2.1.2 Visualização de dados multidimensionais

Segundo Wright (2007) a Visualização Científica apresenta tanto benefícios quanto perigos, pois sempre existem riscos em procedimentos de transformação de dados. Suposições inválidas a respeito dos dados ou sua origem juntamente com a complexidade do sistema visual humano podem levar a representações que podem tanto informar quanto confundir. A projeção, ou produção da imagem representando o conjunto de dados deve levar em consideração os seguintes fatores para melhorar a compreensão: *espaço*, que diz respeito ao posicionamento e tamanho aparente dos objetos que estão presentes; *cores* são parte da maioria das visualizações, podendo confirmar a existência de uma determinada característica nos dados, ou representar a intensidade ou valor para uma segunda variável; *animação* que neste caso corresponde à visualização da variação temporal do conjunto de dados, seja pela sua evolução ou como resultado da iteração do usuário, podendo ser representada pela rápida sucessão de diferentes imagens ou por um conjunto de figuras, cada uma representado um momento no tempo.

Os autores Pao e Meng (1998) abordaram os problemas de se conseguir entender um conjunto de dados multidimensionais e multivariados. A principal ferramenta para viabilizar esta tarefa são os métodos ou algoritmos de redução do número de dimensões dos dados, permitindo assim que eles sejam visualizados em gráficos (projeções) 2D, o que possibilita aos analistas conseguirem entender mais facilmente as relações existentes nos dados.

Ainda segundo Pao e Meng (1998), existem três aspectos na compreensão de dados multidimensionais:

Distribuição dos pontos n-dimensionais: conhecer como os pontos ocupam o espaço de dados (um espaço métrico). A distribuição dos dados é uniforme ou em aglomerados? Segue a mesma distribuição por todo o espaço, ou é regular em uma região e irregular em outra? Assim, a primeira componente do entendimento dos dados é a compreensão a respeito dos relacionamentos entre os pontos, ou vetores, que compõem o conjunto.

Relacionamento funcional: saber se existe um relacionamento funcional entre os valores do campo vetorial do espaço de entrada e o espaço dos valores das propriedades, ou seja, os pontos próximos no espaço de dados correspondem a propriedades similares?

Formação de categorias: criação de aglomerados no espaço de propriedades. Como os pontos no espaço de dados se relacionam com as categorias? Elementos próximos no espaço de dados correspondem à mesma categoria no espaço das propriedades? São estudadas as inconsistências.

O autor Kreuseler (2000) abordou o problema de visualizar uma grande quantidade de dados com relações geográficas e temporais - referentes a dados de observações da vida marinha em um período de 25 anos. O seu objetivo era analisar a relação entre dados hidrológicos (temperatura, salinidade, concentração de oxigênio) e a distribuição de bacalhau em uma determinada região.

Um perfil do terreno foi utilizado para localizar os dados mensurados e gráficos de barra ou de linha foram exibidos em cada local no qual as mensurações foram realizadas. Com esta escolha, não foi possível visualizar uma grande quantidade de dados simultaneamente, forçando o usuário a escolher uma região de interesse e um conjunto de dados a ser analisado, seja de um determinado tempo ou a variação temporal de uma determinada dimensão. Segundo o autor, as seguintes técnicas foram consideradas apropriadas para resolver os problemas apresentados: Coordenadas Paralelas; *Dimensional Stacking* e *Shape Coding*. A aplicação descrita no trabalho foi utilizada para analisar uma base contendo aproximadamente 10.000 registros de informações hidrológicas capturadas em um período de 25 anos. O objetivo principal da aplicação descrita era a visualização de subconjuntos destes dados, selecionados por região, tipo de informação ou evolução temporal.

Santos e Brodlie (2004) realizaram a visualização de dados multidimensionais e multivariados utilizando filtros para selecionar o conjunto de dimensões a serem visualizados, a ferramenta descrita apresenta bons resultados para a navegação em conjuntos extensos de dados, porém a visualização ocorre de forma separada, com um conjunto de dimensões sendo mostrado a cada instante, sendo estas escolhidas pelo usuário da ferramenta.

Os autores Guo, Xiao e Yuan (2011) descreveram um método iterativo para definir a Função de Transferência (FT) utilizada na geração de imagens a partir de dados multidimensionais. A FT deve ser criada com o auxílio da visualização dos dados em um diagrama de coordenadas paralelas e projeções dos *clusters* de dados projetados pelo algoritmo de *Multidimensional Scaling* (MDS).

Em outra linha de pesquisa os autores Lawrence et al. (2011) descreveram um método de visualização de imagens multidimensionais que unifica as diversas bandas de uma imagem multiespectral na banda visível com alto grau de realismo e mantendo as características de distribuição de valores dos dados originais, mostrando exemplos de redução de espaços 4D, 8D e 31D para o espaço 3D com variações nos dados de origem que chegam a 1000:1. “Visualização de dados multidimensionais é de importância crucial para analisar estes conjuntos de dados”. O objetivo do trabalho apresentado é desenvolver um método para mapear uma imagem com um grande número de bandas de valores escalares, em uma imagem com poucas dimensões, ou bandas. De forma simplista, o objetivo deste mapeamento é gerar uma imagem com

poucas dimensões cujos valores escalares sejam consistentes com os valores originais de maneira a preservar (dentro das possibilidades) as distâncias relativas entre as grandezas presentes nos pares de pixels da imagem de entrada.

> "Uma estratégia comum para a análise de conjuntos de dados multi-dimensionais é a projeção destes em um espaço de menor dimensão, no qual o conjunto pode ser visualizado diretamente. Estas técnicas calculam uma projeção linear dos dados, cada uma otimizando um critério diferente. Outro conjunto de técnicas realiza uma transformação não linear nos dados, como no caso dos *Mapas Auto-Organizáveis* (MAO)"[1] (LAWRENCE et al., 2011).

Ainda sobre o trabalho de Lawrence et al. (2011), o problema foi modelado da seguinte maneira: a imagem de entrada com H bandas e N pixels, deve ser mapeada para a imagem de saída com L bandas e N pixels. A matriz de dados de entrada foi denotada por R com dimensões $H \times N$ e a matriz de saída foi denotada por S com dimensões $L \times N$. O objetivo compreendeu solucionar para o conjunto de vetores linha de S, $\{S_1, \cdots, S_N\} \in R^L$, minimizando a Equação (8). A quantidade $\tau(S)$ era referenciada como o valor de STRESS das soluções S

$$\tau(S) = \underbrace{\sum_{i<j} w_{ij}(\delta_{ij}^R - \|S_i - S_j\|)^2}_{\text{Consistência das distâncias}} + \underbrace{\sum_i w_i \|S_i - \bar{S}_i\|^2}_{\text{Restrições de valores}}. \tag{8}$$

O desafio de solucionar esta minimização resulta em uma otimização não linear, devido ao termo $\|S_i - S_j\|$ na restrição de consistência das distâncias. Para resolver o problema, a minimização foi realizada por um processo iterativo conhecido como *majoração*, que resulta em uma sequência de imagens $\{S^k\}$ com valores de stress estritamente decrescentes, $\tau(S^{k-1}) \geq \tau(S^k)$.

Similar ao método de Gauss-Newton, a *majoração* ajusta localmente uma energia quadrática à estimativa atual do stress, e obtém a próxima estimativa ao resolver para o mínimo da energia quadrática. Como a *majoração* da energia é quadrática, minimizar a energia é equivalente a resolver um sistema linear simétrico da forma:

$$(L + \Lambda)S^{K+1} = L^k S^k + \Lambda \bar{S}, \tag{9}$$

onde L é a matriz do Laplaciano, definida pelos pesos das distâncias, Λ é a matriz diagonal com os pesos das restrições e L^k é a matriz Laplaciana que ajusta a relação entre as distâncias desejadas e as distâncias encontradas para solução anterior S^k.

Continuando com o trabalho de Lawrence et al. (2011) e utilizando as definições apresentadas por Gonzalez e Woods (2010), uma imagem multibanda pode ser

[1] Em tradução livre do autor.

representada como uma função f da seguinte forma:

$$f(x,y) = \begin{bmatrix} b_{0,0} & b_{0,1} & \cdots & b_{0,N-1} \\ b_{1,0} & b_{1,1} & \cdots & b_{1,N-1} \\ \vdots & \cdots & \ddots & \vdots \\ b_{M-1,0} & b_{M-1,1} & \cdots & b_{M-1,N-1} \end{bmatrix}, \tag{10}$$

onde b representa um vetor arbitrário no espaço de valores das bandas que compõem a imagem:

$$b_{x,y} = \begin{bmatrix} b_{RGB} \\ b_{IV} \\ b_{UV} \\ b_X \\ \vdots \end{bmatrix}. \tag{11}$$

Em síntese, o trabalho desenvolvido por Lawrence et al. (2011) realizou a transformação do espaço *n*-dimensional das diferentes bandas eletromagnéticas que compõem a imagem original para um espaço tridimensional, para que todas as diferentes informações possam ser representadas em uma imagem no espectro visível, sendo cada uma das dimensões de destino a representação de uma das componentes de cor da imagem resultante.

Um novo método para realizar a projeção de dados multidimensionais foi proposto no trabalho de Peres, Aranha e Pedreira (2013), cujo objetivo era melhor separar as classes de objetos existentes no volume de dados original. Isto foi alcançado ao aplicar a meta heurística *Diferenttial Evolution* para otimizar a divergência das projeções. A medida de divergência estava baseada na Divergência de Cauchy-Schwartz, verificando a separabilidade das classes com a entropia de Renyi e *Information Theoretical Clustering Analysis* (ITC).

2.2 ESPAÇOS E MODELOS DE CORES

Segundo Foley et al. (1996), as sensações visuais causadas pela luz colorida são muito mais ricas do que as provenientes da luz monocromática, e a geração de imagens coloridas exige que as cores sejam especificadas e medidas de forma precisa. Isto normalmente envolve três quantidades conhecidas como matiz, saturação e luminosidade.

Matizes dizem respeito à distinção entre cores, como vermelho, verde e amarelo. Já a saturação representa a distância da cor a um nível de cinza de mesma intensidade, como por exemplo: o vermelho é muito saturado e o rosa pouco saturado ou o azul marinho, que é muito saturado e o ciano que é pouco saturado (FOLEY et

al., 1996). A luminosidade representa a noção acromática da intensidade de percepção de um objeto refletor, e um quarto termo, o brilho, também representa a noção de intensidade, mas para um objeto emissor de luz, como uma lâmpada.

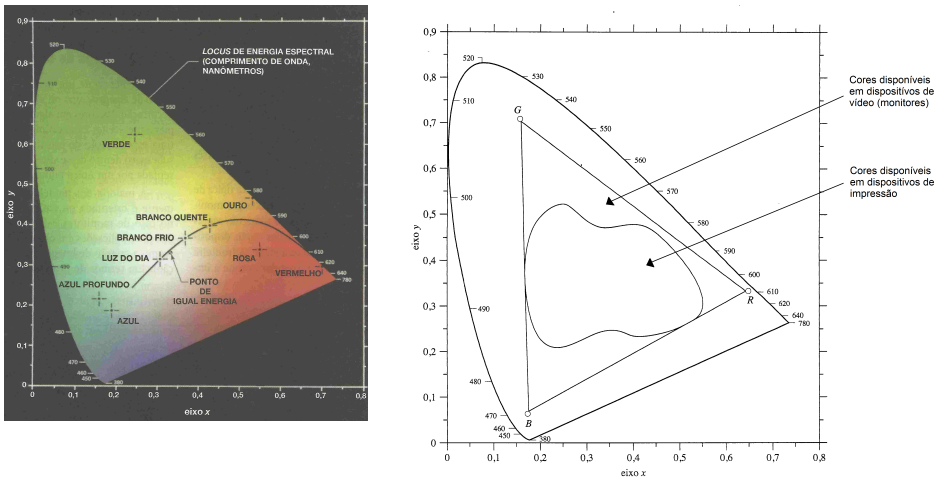
Gonzalez e Woods (2010) explicam que a luz cromática engloba o espectro de energia eletromagnética de aproximadamente 400 nm a 700 nm, e que as células sensíveis às cores do olho humano, os cones, experimentalmente podem ser divididos em três categorias, sendo 65% destes sensíveis à comprimentos de ondas longas, 33% sensíveis à comprimentos de ondas médias e os 2% restantes são sensíveis à comprimentos de ondas curtas, sendo estes últimos os que apresentam maior sensibilidade. Os comprimentos de ondas nas regiões de sensibilidade dos cones correspondem às matizes conhecidas como as cores primárias, que são o vermelho, o verde e o azul. Devido à característica de absorção de energia pelo olho humano, as cores são vistas como uma combinação das cores primárias, porém é importante entender que três comprimentos de ondas específicos não podem, atuando sozinhos, gerar todo o espectro de cores visíveis.

O diagrama de cromaticidade da *Commission Internationale de l'Eclairage* (CIE) representado na Figura 2a mostra o conjunto de cores possíveis de serem retratadas com a composição dos estímulos verde, vermelho e azul. Neste diagrama o *eixo x* corresponde ao percentual de vermelho na composição da cor e o *eixo y* equivale ao percentual de verde. O percentual de azul denotado como a *componente z* da cor é calculado como $z = 1 - (x + y)$. Este diagrama é útil para a determinação de todas as cores que podem ser obtidas da combinação aditiva entre quaisquer duas cores no diagrama, através de um segmento de reta que as une no diagrama e passa por todas as resultantes possíveis. Este conceito pode ser estendido à combinação de três cores, formando assim uma região triangular no interior do diagrama que delimita todas as cores que podem ser obtidas a partir da adição proporcional das três componentes iniciais.

O fato da impossibilidade de representação das cores por um conjunto de três matizes pode ser melhor compreendido na apresentação do conjunto de cores que podem ser exibidas em um monitor colorido (que utiliza as componentes vermelho, verde e azul), representado pelo triângulo na Figura 2b, evidenciando que existe um conjunto de cores que não é possível representar no monitor, até os limites definidos no diagrama de cromaticidade da CIE. A região, que está delimitada no interior do triângulo, corresponde às cores que podem ser representadas em um dispositivo de impressão de alta qualidade, deixando claro também que o conjunto de cores disponíveis neste último é ainda menor, exigindo que sejam realizadas transformações nos conjuntos de cores para obter o mesmo efeito visual em diferentes dispositivos (GONZALEZ; WOODS, 2010).

FIGURA 2 – (a) Diagrama de cromaticidade da CIE (b) Gama de cores típico para monitores e dispositivos de impressão

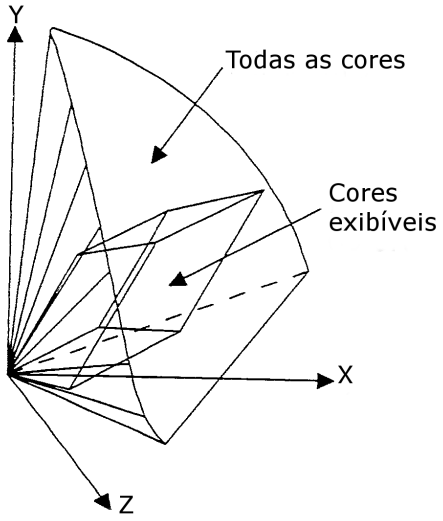
(a) (b)



FONTE: Adaptado de Gonzalez e Woods (2010, p. 263, 264)

Em dispositivos utilizados como saída para processos de visualização, um modelo de cores é especificado por um sistema de coordenadas tridimensional juntamente com um subconjunto visível no sistema de coordenadas no qual todas as cores são representadas. Esta relação pode ser melhor compreendida na representação exibida na Figura 3, do volume de cores do diagrama de cromaticidade da CIE, no qual o sistema de coordenadas *Red, Green, Blue* (RGB) de um monitor está inscrito (FOLEY et al., 1996).

FIGURA 3 – Representação do espaço de cores RGB no espaço de cores da CIE



FONTE: Adaptado de Foley et al. (1996, p. 585)

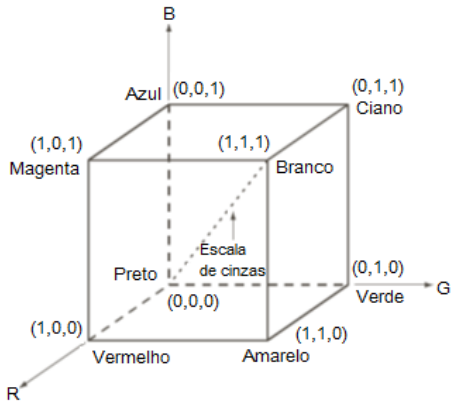
Para Gonzalez e Woods (2010) um modelo de cores, também chamado de espaço de cores ou sistema de cores, é uma forma padronizada de especificar as cores. Sendo o modelo RGB utilizado em monitores e o modelo *Cyan, Magenta, Yel-*

low (CMY) ou *Cyan, Magenta, Yellow, Black* (CMYK) utilizado em dispositivos de impressão, e de acordo com o conteúdo da Figura 2b, estes sistemas não conseguem representar o mesmo conjunto de cores. Além destes, um outro conjunto de modelos foi desenvolvido para levar em conta as noções intuitivas de matiz, saturação e brilho, dentre eles o *Hue, Saturation, Value* (HSV), o *Hue, L, Saturation* (HLS) e o *Hue, Value, C* (HVC) (FOLEY et al., 1996).

2.2.1 Modelo de cores RGB

No modelo RGB, as cores são representadas por suas componentes espectrais primárias, o vermelho (*Red*), o verde (*Green*) e o azul (*Blue*), onde o sistema de coordenadas cartesiano normalizado é utilizado para representar os valores de cada componente, como ilustrado na Figura 4 (GONZALEZ; WOODS, 2010).

FIGURA 4 – Esquema do cubo de cores do sistema RGB



FONTE: Adaptado de Gonzalez e Woods (2010, p. 265)

Este modelo é utilizado em dispositivos cuja característica de composição de cores é aditiva, por exemplo, para obter uma das cores secundárias, o ciano, o magenta e o amarelo, duas das cores primárias são adicionadas uma à outra, o que é evidenciado nos vértices do cubo da Figura 4. Assim, o amarelo é formado pela soma do vermelho com o verde; o Magenta representa a soma do vermelho com o azul; e o ciano é o resultado da soma entre o verde e o azul. A origem do sistema de coordenadas representa a ausência de cores, o preto, e a coordenada $(1, 1, 1)$ representa a soma de todas as cores primárias, resultando na luz branca, e todas as cores possíveis de serem representadas são obtidas por uma combinação destas componentes. Na Figura 4 percebe-se que a diagonal principal do espaço de cores RGB representa as cores acromáticas, também conhecidas com escala de cinzas, e na Figura 5 as cores do espaço estão representadas na mesma projeção utilizada no esquema da Figura 4 (FOLEY et al., 1996; GONZALEZ; WOODS, 2010).

FIGURA 5 – Cubo de cores do sistema RGB

FONTE: Gonzalez e Woods (2010, p. 266)

2.2.2 Modelo de cores CMY

As cores ciano, magenta e amarelo (CMY), conhecidas como primitivas subtrativas ou também como cores primárias de pigmentos, são as cores complementares ao vermelho, verde e azul, respectivamente, e são obtidas ao serem aplicados filtros ou processos que subtraem cores da luz branca, portanto o modelo CMY é baseado em um sistema subtrativo de formação de cores, ou seja, para uma superfície pigmentada com a cor ciano e iluminada com a luz branca, nenhuma cor vermelha é refletida, assim o ciano subtrai o vermelho da luz branca (no modelo RGB o ciano é formado pela soma do verde o do azul, sem a adição da componente vermelha) (GONZALEZ; WOODS, 2010).

Como os modelos CMY e RGB são complementares, as relações entre os valores das cores de cada sistema podem ser expressas utilizando a Equação (12).

$$\begin{bmatrix} C \\ M \\ Y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{12}$$

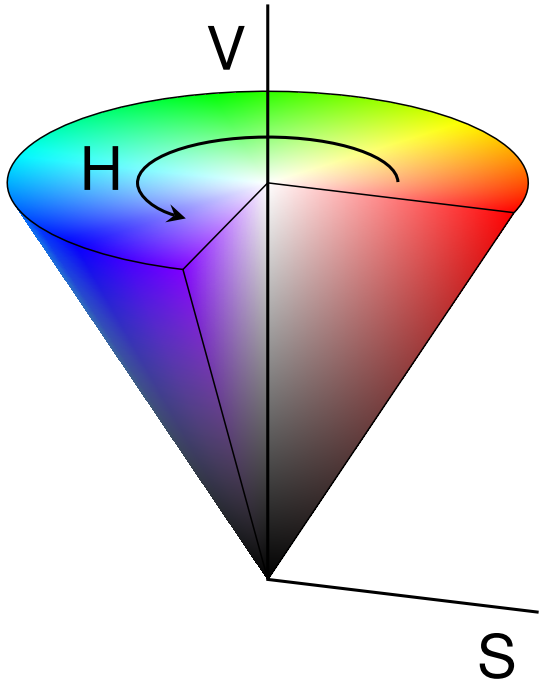
Segundo Gonzalez e Woods (2010), a mesma quantidade de pigmentos ciano, magenta e amarela não resultam na cor preta, então em sistemas de impressão, normalmente é adicionada uma quarta cor, o preto, formando então o modelo CMYK, no qual o preto é utilizado para substituir quantidades iguais dos pigmentos primários, definidos no padrão CMY, seguindo as relações apresentadas na Equação (13) (FOLEY et al., 1996).

$$
\begin{aligned}
K &= \min(C_{CMY}, M_{CMY}, Y_{CMY}), \\
C_{CMYK} &= C_{CMY} - K, \\
M_{CMYK} &= M_{CMY} - K, \\
Y_{CMYK} &= Y_{CMY} - K.
\end{aligned} \tag{13}
$$

2.2.3 Modelo de cores HSV

Tanto o modelo RGB quanto o CMY ou CMYK são adequados para utilização em equipamentos de formação de imagens, como monitores e impressoras, porém não são adequados para a descrição de cores para a interação humana, ou seja, ao observar um objeto, uma pessoa descreve a sua cor em termos de matiz, saturação e brilho - *Hue, Saturation, Brightness* (HSB), que também pode ser denominado valor - *Hue, Saturation, Value* (HSV), ou intensidade - *Hue, Saturation, Intensity* (HSI). Neste sistema uma cor pura é representada por uma matiz, por exemplo o vermelho, o verde, o amarelo, e assim por diante. Já a saturação, indica o grau de diluição da cor pura por luz branca e, por último, o brilho representa uma noção subjetiva de intensidade acromática. Segundo Foley et al. (1996), o sistema de coordenadas deste modelo é cilíndrico, no qual as cores são representadas em um subconjunto definido por um cone, de acordo com o conteúdo da Figura 6 (GONZALEZ; WOODS, 2010).

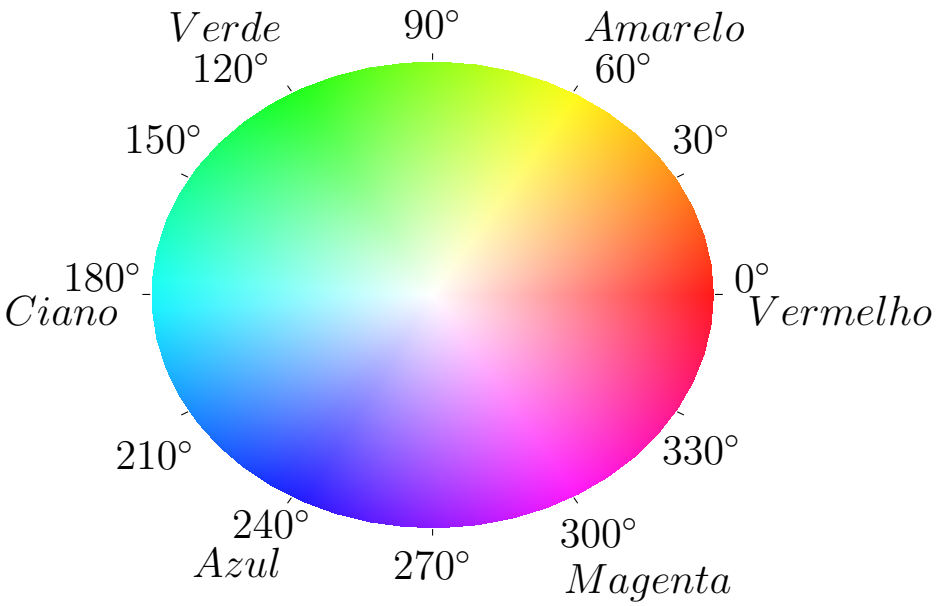
FIGURA 6 – Representação do espaço de cores HSV



FONTE: Autoria própria

Na Figura 6 as matizes estão dispostas em volta do eixo vertical, sendo o ângulo de deslocamento a unidade de medida utilizada, com o vermelho posicionado em 0°, o verde em 120° e o azul em 240°. As cores complementares aparecem separadas por 180°, ficando posicionadas diametralmente opostas uma da outra no plano superior do modelo, representado na Figura 7. A saturação de uma cor varia no eixo central de 0 até 1, nas bordas do hexágono, indicando quão diluída ou pura em relação à luz branca está a matiz.

FIGURA 7 – O plano de cores do modelo HSV



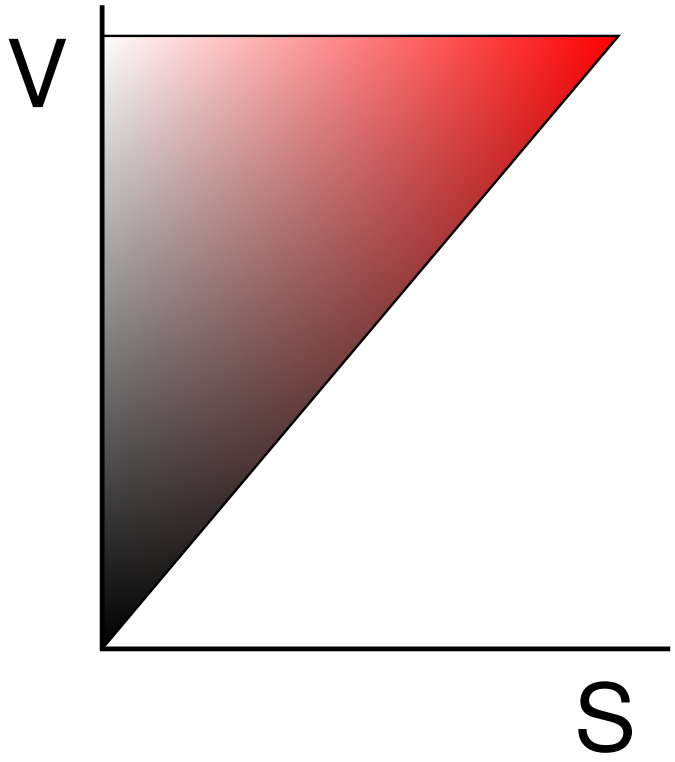
FONTE: Autoria própria

O brilho, ou intensidade, varia de 0 para o preto, até 1, para o branco, portanto quando o valor da saturação for 0 ($S = 0$), a matiz H é irrelevante e a variação de V produz a escala de cinzas, e quando V tiver valor 0 ($B = 0$), tanto a matiz quanto a saturação são irrelevantes. Na Figura 8 está representada a matiz correspondente ao vermelho, evidenciando as variações dos valores de saturação e brilho.

Existe uma relação entre os espaços de cores, sendo possível realizar a conversão das representações de um para outro. Neste sentido, para a conversão dos valores do sistema RGB para o sistema HSV são utilizadas as relações apresentadas nas Equações (14) e (15).

Para realizar a conversão de uma cor no modelo RGB são necessários os se-

FIGURA 8 – Relação entre Saturação e Brilho no diagrama de cores HSV



FONTE: Autoria própria

guintes valores intermediários:

$$
\begin{aligned}
M &= \max(R, G, B) \\
m &= \min(R, G, B) \\
C &= M - m \\
H' &= \begin{cases} \text{indefinido}, & se\ C = 0 \\ \frac{G-B}{C}, & se\ M = R \\ \frac{B-R}{C} + 2, & se\ M = G \\ \frac{R-G}{C} + 4, & se\ M = B \end{cases},
\end{aligned}
\tag{14}
$$

e a partir destes, calcula-se os valores dos componentes H, S e V para representar a cor.

$$
\begin{aligned}
H &= 60° \times H' \\
S &= \begin{cases} 0, & se\ M = 0 \\ \frac{C}{M}, \end{cases} \\
V &= M
\end{aligned}
\tag{15}
$$

Da mesma maneira, dada uma cor no sistema HSV, a sua conversão para o sistema RGB segue as relações apresentadas nas Equações (16) e (17).

Igualmente, esta conversão necessita dos seguintes valores intermediários:

$$
\begin{aligned}
C &= V \times S \\
H' &= \frac{H}{60°} \\
X &= C\,(1 - |(H'\ mod\ 2) - 1|) \\
m &= V - C \\
(R_1, G_1, B_1) &= \begin{cases}
(0,0,0) & se\ H \text{ for indefinido.} \\
(C,X,0) & se\ 0 \leq H' < 1. \\
(X,C,0) & se\ 1 \leq H' < 2. \\
(0,C,X) & se\ 2 \leq H' < 3. \\
(0,X,C) & se\ 3 \leq H' < 4. \\
(X,0,C) & se\ 4 \leq H' < 5. \\
(C,0,X) & se\ 5 \leq H' < 6.
\end{cases}
\end{aligned}
\tag{16}
$$

A partir destes valores, as coordenadas no sistema RGB podem ser calculadas para a cor desejada da seguinte maneira:

$$(R, G, B) = (R_1 + m, G_1 + m, B_1 + m) \quad . \tag{17}$$

2.3 ANÁLISE DE DADOS MULTIDIMENSIONAIS

Os autores Baumgartner e Somorjai (2001) apresentaram como uma definição o fato que a visualização de dados multidimensionais faz parte da estatística computacional e da análise exploratória de dados - *Exploratory Data Analysis* (EDA) - e também que a visualização do conjunto completo de dados antes da redução de características (redução de dimensões) é importante na exploração dos dados, e da mesma forma para métodos inferenciais. Segundo Tukey (1977 apud BAUMGARTNER; SOMORJAI, 2001) a visualização ou projeção de dados multidimensionais deve ser parte integrante da EDA.

Em seu artigo, Baumgartner e Somorjai (2001) apresentaram um método para formar imagens a partir do coeficiente de correlação de agrupamentos de dados, seguindo seu posicionamento ou centroide em instantes de tempo, ou *frames* temporais. Cada *frame* temporal forma uma coluna na imagem de saída, e a coordenada vertical representa o coeficiente de correlação do centroide do *cluster* com os demais *frames* temporais. Os dados são dispostos de acordo com a *Minimum Spanning Tree* (MST), sequências temporais (*time-course* - (TC)). Os nós da MST são as posições do *cluster*, e as arestas são calculadas como as distâncias euclidianas deste *cluster* em cada *frame* temporal.

Os autores Yuan et al. (2013) descreveram um sistema para analisar um conjunto de dados multidimensional que constrói uma árvore de subespaços a partir de interações do usuário. O objetivo do sistema referido era fornecer aos especialistas um método visual para auxiliar na descoberta de relações e grupos no conjunto de dados, e de dimensões, que não são facilmente visíveis em outras formas de análise visual de dados multidimensionais. Como ferramentas foram utilizados, principalmente, *Principal Component Analysis* (PCA) e MDS para realizar a projeção dos dados em gráficos 2D e o cálculo do coeficiente de correlação de Pearson, para realizar a representação gráfica das dimensões dos dados.

Santos e Brodlie (2004) expuseram um método para realizar a visualização de dados multidimensionais e multivariados utilizando filtros para selecionar o conjunto de dimensões a serem visualizadas. O autores apresentaram bons resultados para a navegação em conjuntos extensos de dados, porém a visualização ocorre de forma separada, com um conjunto de dimensões, escolhidas pelo usuário da ferramenta, exibido a cada instante.

Segundo Tsai (2012), ao se realizar a redução do número de dimensões de um conjunto de dados, espera-se encontrar estruturas ocultas que auxiliem na compreensão, bem como na visualização do conjunto. Técnicas como a PCA e a MDS são capazes de trabalhar com dados que sejam lineares por natureza. Trabalhos recentes descrevem técnicas para trabalhar conjuntos de dados que obedecem a certas condições de não-linearidade, são elas: *Locally Linear Embedding* (LLE); *Hessian Locally Linear Embedding* (HLLE); *Isometric Feature Mapping* (Isomap); *Local Tangent Space Alignment* (LTSA); *Kernel Principal Component Analysis*; *Diffusion Maps* e *Multilayer Autoencoders*. Ao se realizar a redução de dimensões é necessário avaliar a qualidade do resultado obtido, e para isso a autora compara algumas medidas utilizadas para julgar a qualidade de resultados de baixa dimensionalidade e propõe um método que utiliza a correlação das distâncias de pares geodésicos, que apresenta melhores resultados caso os dados sejam organizados não-linearmente.

2.4 REDUÇÃO DO NÚMERO DE DIMENSÕES

Nesta seção são apresentados trabalhos que descrevem técnicas e algoritmos que realizam a redução do número de dimensões de um conjunto de dados, com os objetivos de facilitar a análise e a visualização destes conjuntos, tanto de forma automatizada como por parte de especialistas.

Segundo Pao e Meng (1998), “o impedimento para entender grandes conjuntos de dados multivariados não recai sobre a dificuldade computacional de calcular várias distribuições e relacionamentos, mas sim na falta de meios convenientes para reter tal

informação de uma forma que possa ser utilizada intuitiva e convenientemente."[2]

O procedimento de Redução Dimensional (RD) consiste em mapear os elementos de um conjunto com *n* dimensões para uma representação que mantenha, da melhor forma possível, as relações entre elementos e seus agrupamentos em um conjunto com m dimensões, sendo $m \ll n$ (BORG; GROENEN, 2005; van der MAATEN; POSTMA; van den HERIK, 2009; ADHIANTO et al., 2013). Portanto, para um conjunto de h elementos $X^n = \{x_i \in \mathbb{R}^n\}_{1 \leq i \leq h}$, um algoritmo de RD pode ser interpretado como uma função

$$f : \mathbb{R}^n \times T \rightarrow \mathbb{R}^m, \tag{18}$$

que mapeia, por meio de uma transformação T, cada um dos x_i elementos em um novo elemento y_i no espaço $\mathbb{R}^m$ (MARTINS et al., 2014).

Um conjunto multidimensional, após a realização da redução de dimensões, deve manter as relações de vizinhança entre os vetores, ou seja, um conjunto de pontos próximos no espaço *n*-dimensional deve também formar um conjunto de vizinhos na projeção dos dados no espaço m-dimensional. Cada um dos métodos de RD apresentam uma peculiaridade com relação à disposição da vizinhança dos pontos projetados (MARTINS et al., 2014).

Existem dezenas de métodos utilizados para realizar a RD apresentados na literatura, classificados de acordo com o algoritmo utilizado para calcular a função f, apresentada na Equação (18). Os métodos de RD também podem ser classificados quanto ao tipo de transformação realizada. Assim, os que se baseiam em transformações lineares do produto interno são definidos como técnicas de projeção. Já os métodos que são capazes de determinar as relações de distância em uma estrutura de dados não-linear são definidos como métodos baseados em aprendizado de topologia[3] (ENGEL; HÜTTENBERGER; HAMANN, 2012).

Alguns dos métodos mais utilizados são o MDS, o PCA, o Isomap, o LLE, o *Stochastic Neighbor Embedding* (SNE), o *Stochastic Proximity Embedding* (SPE) e diversas configurações ou modelos de Redes Neurais Artificiais. Esta lista não tem a pretensão de ser completa ou qualificar os métodos, mas sim apresentar um conjunto dos mais citados, de maneira a expor a diversidade de abordagens ao problema de RD.

Os autores Engel, Hüttenberger e Hamann (2012) apresentam um estudo com a descrição de um conjunto de métodos para realizar a RD utilizados na análise e visualização de dados multidimensionais. Descrevem e classificam o *Principal Components Analysis*, o *Metric Multidimensional Scaling*, também conhecido como *Classical Multidimensional Scaling*, o *Kernel Principal Components Analysis*, o *Non-metric*

[2] Em tradução livre do autor.
[3] Manifold learning

Multidimensional Scaling, o *Isometric Feature Mapping*, o *Locally Linear Embedding*, o *Piecewise Laplacian-based Pjection* e o *Multigrid Multidimensional Scaling*.

No trabalho de Pao e Meng (1998), são citados sete métodos utilizados para a redução das dimensões dos dados relacionados com trabalhos no campo da inteligência artificial, são eles: *Ratio-Conserving Map*; Transformada de Karhunen-Loeve; Mapa de características de Kohonen; *Generative Topographic Mapping*; *Autoassociative Mapping Approach*; *Nonlinear Variance Conserving Approach* e *Equalized Orthogonal Mapping*.

Uma nova abordagem para a visualização de dados multidimensionais utilizando *Self-Organizing Maps* (SOMs), ou Mapas de Kohonen, foi apresentada no trabalho de Shieh e Liao (2012). De acordo com os autores, estes mapas podem ser utilizados para projetar dados multidimensionais em mapas topológicos de baixa dimensionalidade. Foram citadas versões aprimoradas dos algoritmos, o *Visualization induced SOM* (ViSOM) e o *Probabilistic Regularized SOM* (PRSOM) capazes de aprimorar as características de visualização e manter as distâncias entre neurônios, com melhores resultados que os algoritmos de *Curvilinear Component Analysis* (CCA), MDS e *Sammon's Mapping*. O método proposto se distinguiu por permitir trabalhar com dados não rotulados e apresentar graficamente as distâncias entre neurônios de acordo com uma escala de cores e a probabilidade mensurável de cada neurônio.

A seguir são apresentadas as descrições para alguns destes métodos, conjuntamente com a descrição do sistema de Coordenadas Estrela, uma solução para análise de dados que também pode ser utilizada para redução dimensional. Além destas, uma seção expõe considerações a respeito da qualidade das técnicas de RD.

2.4.1 *Multidimensional Scaling*

Segundo Bae, Qiu e Fox (2012), *Multidimensional Scaling* é um termo geral para designar as técnicas de redução da dimensão de dados baseadas na informação de proximidade entre pares de pontos ou elementos do conjunto de dados. Borg e Groenen (2005) definem o MDS como uma tarefa cujo objetivo é encontrar uma configuração de pontos que representam objetos, em um espaço de baixa dimensionalidade, de tal modo que as distâncias entre quaisquer dois pontos correspondam às dissimilaridades dos objetos que representam, o mais fielmente possível.

O MDS é um método que representa medidas de similaridade (ou dissimilaridade) entre pares de objetos como distâncias entre pontos em um espaço com poucas dimensões. A representação gráfica do MDS permite aos analistas, literalmente, olhar para os dados e explorar a sua estrutura visualmente (BORG; GROENEN, 2005; TSAI, 2012). O método MDS define uma aproximação aceitável como sendo a que captura as relações entre as distâncias dos elementos de uma forma ótima, mais pre-

cisamente, as relações dos produto interno entre eles, assumindo que a medida de distância utilizada seja métrica. Caso seja utilizada a distância Euclidiana, o MDS é equivalente ao *Principal Component Analysis* (descrito na Seção 2.4.4), incluindo as transformações de escala e rotação no espaço de origem (ENGEL; HÜTTENBERGER; HAMANN, 2012).

A informação de proximidade dos pares de dados é dada por uma matriz de ordem $N \times N$ ($\Delta = [\delta_{ij}]$), onde N é o número de pontos e δ_{ij} é o valor de dissimilaridade entre os pontos i e j no espaço original. A matriz Δ é simétrica, $\delta_{ij} = \delta_{ji}$, não negativa, $\delta_{ij} > 0$ e possui diagonal principal igual a 0, $\delta_{ii} = 0$.

O resultado do MDS pode ser representado por uma matriz X de dimensões $N \times L$, onde L é o número de dimensões do espaço de destino, e cada ponto $x_i \in \mathbb{R}^L$ $(i = 1, \ldots, N)$ reside na i-ésima linha de X.

Segundo Borg e Groenen (2005) o MDS consiste em, a partir de uma matriz X de elementos no espaço *n*-dimensional, calcular uma matriz com os quadrados das dissimilaridades destes elementos, Δ^2, para em seguida aplicar a operação denominada centralização dupla (*double centering*), que consiste em calcular a matriz B_Δ, dada por:

$$B_\Delta = -\frac{1}{2}\, J\, \Delta^2\, J, \tag{19}$$

sendo J a matriz de centralização dada por: $J = I - n^{-1} U$, onde I é a matriz identidade, U é uma matriz cujos elementos são iguais a 1 e *n* é o número de dimensões do conjunto de elementos.

A matriz Δ^2 deve ser decomposta em seus autovalores Λ, e autovetores Q, de modo que:

$$B_\Delta = Q\, \Lambda\, Q' = (Q\Lambda^{1/2})(Q\Lambda^{1/2})' = YY'. \tag{20}$$

Após a decomposição, considera-se a matriz formada pelos primeiros m autovalores maiores que zero para compor a matriz Λ_+ e Q_+ é a matriz formada pelas primeiras m colunas de Q, fazendo com que a matriz de coordenadas resultante seja:

$$Y = Q_+ \Lambda_+^{1/2}. \tag{21}$$

Este método minimiza a função de perda dada por:

$$L(Y) = \|YY' - B_\Delta\|^2. \tag{22}$$

A avaliação da relação entre os dados dos espaços de origem e destino é feita utilizando-se as Equações: STRESS (23) e SSTRESS (24):

$$\sigma(X) = \sum_{i<j\leq N} w_{ij}\, (\delta_{ij} - d_{ij}(X))^2 \tag{23}$$

e

$$\sigma^2(X) = \sum_{i<j\leq N} w_{ij} \left[(\delta_{ij})^2 - (d_{ij}(X))^2 \right]^2, \tag{24}$$

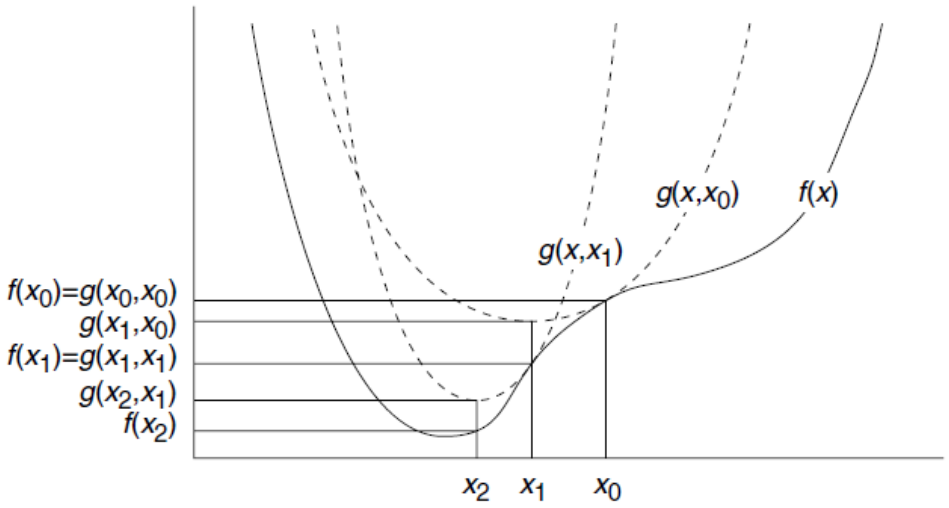
onde w é uma matriz de pesos, δ_{ij} são os elementos da matriz de dissimilaridades e d_{ij} são os elementos da matriz de distâncias entre os componentes Y, que é a matriz de projeções dos elementos de X.

2.4.2 *Scaling by Mojorizing a Complicated Function*

Conforme descrito no trabalho de Borg e Groenen (2005), o algoritmo *Scaling by Mojorizing a Complicated Function* (SMACOF) resolve o MDS por um processo iterativo. Ele minimiza o valor da função STRESS apresentada na Equação (24), a partir de uma matriz X de elementos no espaço *n*-dimensional e da matriz Δ que é formada pelas dissimilaridades destes elementos. A função STRESS representa as diferenças entre as medidas das dissimilaridades representadas na matriz Δ e os valores de distância entre as projeções dos elementos de X no espaço *m*-dimensional de destino.

O objetivo do MDS é obter a matriz X com o menor valor de STRESS possível. Para isso um processo iterativo é aplicado no qual uma sequência de valores monotonicamente não crescentes é gerada, este procedimento é chamado de majoração iterativa e está ilustrado na Figura 9, onde se pode perceber que os pontos mínimos da série de funções $g(x_i, x_j)$ se aproximam do ponto mínimo da função $f(x)$, que é o objetivo do processo (BORG; GROENEN, 2005).

FIGURA 9 – Iterações do processo iterativo de majoração



FONTE: Borg e Groenen (2005, pág. 180)

O algoritmo de majoração do valor da função de STRESS é o resultado do trabalho de De Leeuw, em 1977, e recebeu o nome de SMACOF e, segundo (BORG;

GROENEN, 2005), outros algoritmos foram desenvolvidos com o mesmo objetivo, porém o SMACOF é mais simples e poderoso porque garante a convergência do valor da função de STRESS.

De acordo com o exposto por Borg e Groenen (2005), a expressão de STRESS (23) pode ser escrita como:

$$\begin{aligned}\sigma(X) =& \sum_{i<j} w_{ij} \left(\delta_{ij} - d_{ij}(X)\right)^2 \\ =& \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij} d_{ij}^2(X) - 2\sum_{i<j} w_{ij}\delta_{ij} d_{ij}(X) \\ =& \eta_{ij}^2 + \eta^2(X) - 2\rho(X)\ . \end{aligned} \tag{25}$$

O termo η_{ij}^2 não depende dos valores de X e, portanto, é constante. Considerando-se e_i como a i-ésima coluna da matriz identidade, pode-se escrever a distância entre dois elementos como:

$$\begin{aligned} d_{ij}^2 =& \sum_{a=1}^{m} x_a^{'}(e_i - e_j)(e_i - e_j)^{'} x_a \\ =& tr\ X^{'} A_{ij} X, \end{aligned} \tag{26}$$

portanto:

$$\begin{aligned} w_{ij} d_{ij}^2(X) =& w_{ij} tr\ X^{'} A_{ij} X \\ =& tr\ X^{'}(w_{ij} A_{ij}) X \end{aligned} \tag{27}$$

e

$$\begin{aligned} \eta^2(X) = \sum_{i<j} w_{ij} d_{ij}^2(X) =& tr\ X^{'} \left(\sum_{i<j} w_{ij} A_{ij}\right) X \\ =& tr\ X^{'} V X\ . \end{aligned} \tag{28}$$

Para o termo $-2\rho(X)$ pode-se escrever[4]:

$$\begin{aligned} -\rho(X) =& -\sum_{i<j} \left(w_{ij}\delta_{ij}\right) d_{ij}(X) \\ \leq& -tr\ X^{'} \left(\sum_{i<j} b_{ij} A_{ij}\right) Z \\ =& -tr\ X^{'} B(Z) Z\ , \end{aligned} \tag{29}$$

[4] Os passos adicionais para a obtenção da equação de majoração da função de STRESS podem ser observados no texto de Borg e Groenen (2005).

onde a matriz $B(Z)$ é definida como:

$$b_{ij} = \begin{cases} -\frac{w_{ij}\delta_{ij}}{d_{ij}(Z)} & \text{para } i \neq j \text{ e } d_{ij}(Z) \neq 0 \\ 0 & \text{para } i \neq j \text{ e } d_{ij}(Z) = 0 \end{cases} \tag{30}$$

$$b_{ii} = -\sum_{j=1;j\neq i}^{n} b_{ij} \ . \tag{31}$$

Caso ocorra a igualdade $Z = X$, obtém-se:

$$-\rho(x) = -tr\ X^{'}B(X)X \leq -tr\ X^{'}B(Z)Z \ , \tag{32}$$

que é a desigualdade de majoração e, portanto, $\rho(X)$ pode ser majorado pela função $-tr\ X^{'}B(Z)Z$.

Baseando-se nas definições apresentadas acima, a função de STRESS pode ser reescrita como:

$$\begin{aligned} \sigma_r(X) =& \eta_\delta^2 + tr\ X^{'}VX - 2\,tr\ X^{'}B(X)X \\ \leq& \eta_\delta^2 + tr\ X^{'}VX - 2\,tr\ X^{'}B(Z)Z = \tau(X,Z) \ , \end{aligned} \tag{33}$$

sendo $\tau(X, Z)$ a função de majoração, que pode ser minimizada igualando-se sua derivada a zero:

$$\nabla\tau(X,Z) = 2VX - 2B(Z)Z = 0 \ . \tag{34}$$

A solução do sistema de equações resultante leva à formula de atualização do algoritmo SMACOF:

$$X^u = V^+B(Z)Z \ . \tag{35}$$

Para o caso especial em que todos os pesos sejam iguais a 1, $w_{ij} = 1$, a fórmula de atualização do algoritmo é simplificada para:

$$X^u = n^{-1}B(Z)Z \ . \tag{36}$$

Resumidamente, o algoritmo consiste em, a partir de uma projeção inicial Y não aleatória, calcular as diferenças entre as distâncias utilizando a função STRESS, e, enquanto o seu valor for maior que um limite de precisão, ou um máximo de iterações não for atingido, atualizar a matriz Y utilizando $Y^u = n^{-1}B(Y)Y$ no caso da matriz de pesos ter todos os elementos iguais a 1, ou $Y^u = V^+B(Y)Y$ caso contrário, sendo a matriz V^+ a inversa das somas ponderadas das distâncias entre elementos de Y, e $B(Y)$ a matriz formada pela razão ponderada entre as dissimilaridades dos elementos de X e Y.

A limitação da maioria das aplicações do método MDS é a necessidade de $O(N^2)$ espaços de memória e $O(N^2)$ operações ou computações, portanto pode ser considerado um problema limitado por restrições de memória (BAE; QIU; FOX, 2012).

2.4.3 *Stochastic Proximity Embedding*

Este método está descrito no trabalho de Najim e Lim (2014) como não linear e iterativo, que consiste em atualizar as projeções de cada elemento considerando as suas distâncias aos demais componentes do conjunto que sejam menores que um raio r_C.

A partir de uma projeção aleatória inicial, a solução é refinada realizando-se o sorteio de um ponto i e utilizado-o para ajustar as coordenadas de todos os demais elementos da projeção, com $j \neq i$, utilizando a seguinte regra:

$$y_j = y_j + \lambda(t_k)\, S(\delta_{ij})\, \frac{\delta_{ij} - d_{ij}}{d_{ij} + \epsilon}\, (y_j - y_i) \tag{37}$$

$$S(\delta_{ij}) = \begin{cases} 1 & se(\, (\delta_{ij} \leq r_C) \vee (\, (\delta_{ij} > r_C) \wedge (d_{ij} < \delta_{ij})\,)\,) \\ 0 & \text{caso contrário}, \end{cases} \tag{38}$$

onde x_i e x_j representam as coordenadas dos elementos i e j, $\lambda(t_k)$ é a taxa de aprendizado, ϵ é uma constante para evitar divisão por 0, δ_{ij} representa a distância entre os elementos no espaço *n*-dimensional e d_{ij} a distâncias de suas projeções no espaço *m*-dimensional.

2.4.4 *Principal Component Analysis*

O PCA é descrito por van der Maaten, Postma e van den Herik (2009) como sendo, matematicamente, equivalente ao método *Classical Scaling*, pois em ambos os métodos o objetivo é minimizar a função de perda, ou seja, busca-se uma representação para os dados na qual as diferenças de valores para uma determinada medida de distância, aplicada entre os pares de elementos, seja a menor possível. Para o MDS, utiliza-se a distância euclidiana e no PCA é utilizada a matriz de covariâncias dos elementos.

Segundo Engel, Hüttenberger e Hamann (2012) o PCA é um dos primeiros métodos de RD descritos na literatura, no qual os dados são projetados em um subespaço linear, no qual os dados projetados possuem máxima variância, sendo que esta projeção não distorce os dados.

Da mesma forma que no MDS, o PCA resolve uma decomposição em autovalores como $cov(X)M = \lambda M$, na qual M representa a matriz que mapeia os elementos do espaço *n*-dimensional para o espaço *m*-dimensional, e λ é a matriz formada pelos autovalores de $cov(X)$.

O ponto fraco do método, devido a sua natureza linear, é sua incapacidade de tratar corretamento dados com características não lineares (ENGEL; HÜTTENBERGER; HAMANN, 2012).

2.4.5 *Isometric Feature Mapping*

O Isomap é um método de RD proposto por Tenenbaum, Silva e Langford (2000), que trabalha com as distâncias geodésicas, e não com as distâncias euclidianas, entre os elementos do conjunto de pontos no espaço *n*-dimensional. O seu objetivo é capturar a geometria dos pontos no conjunto *n*-dimensional e mantê-la na projeção.

A redução dimensional é realizada utilizando-se o MDS, porém para trabalhar com as distâncias geodésicas, para cada elemento, são calculados os seus k vizinhos mais próximos, formando assim um grafo conectado, e a partir deste, o menor caminho entre cada par de elementos corresponde à sua distância, que pode ser calculada por algoritmos como o caminho mais curto de Djikstra ou Floyd (van der MAATEN; POSTMA; van den HERIK, 2009; ENGEL; HÜTTENBERGER; HAMANN, 2012).

Um problema associado ao Isomap é que após a centralização dupla[5] das distâncias geodésicas, a matriz B_{Δ} resultante pode não ser semi-definida positiva, o que não garante a existência de seus autovetores e autovalores. O método *Maximum Variance Unfolding* (MVU) garante esta situação partindo da ideia de desdobrar as superfícies topológicas impondo a restrição de que as distâncias locais entre os pontos sejam mantidas, o que é otimizado com relação à máxima variância (ENGEL; HÜTTENBERGER; HAMANN, 2012).

2.4.6 *Ratio-Conserving Map*

Trata-se de uma técnica de redução das dimensões dos dados na qual uma rede *Multi Layer Perceptron* (MLP) é treinada, sem os valores objetivos de treinamento, com um conjunto de dados multidimensionais, com uma saída para um espaço de menor dimensão (por exemplo o espaço 2D). O treinamento é realizado levando em consideração as distâncias euclidianas dos dados, ou seja, para cada tupla no conjunto de entrada são calculadas as distâncias para as demais tuplas, o mesmo é realizado para os valores de saída. Caso as distâncias sejam diferentes, um erro é calculado com base nas diferenças das distâncias entre os dois conjuntos, e aplicado como fator de correção na rede neural. Isto faz com que as distâncias entre os dados de entrada e os de saída permaneçam aproximadamente iguais, tornando possível analisar os relacionamentos entre as tuplas de forma visual em um gráfico 2D (PAO; MENG, 1998).

2.4.7 Transformada de Karhunen-Loeve

É um procedimento para descobrir padrões que são combinações lineares dos padrões originais de entrada para o método. Os novos padrões, cujos valores não se

[5] Definida na Seção 2.4.1

distinguem muito uns dos outros, podem ser descartados durante os procedimentos de classificação (PAO; MENG, 1998).

2.4.8 *Nonlinear variance conserving approach*

Nonlinear variance conserving approach (NLVC) ou "Conservação não linear da variância"[6] é uma extensão da Transformada de Karhunen-Loeve na qual a transformação não linear é implementada por uma rede neural multi camadas, cujo critério de treinamento exige que a variância total das saídas seja uma fração da variância da entrada (PAO; MENG, 1998).

2.4.9 Mapa de características de Kohonen

Ou mapa auto organizável, é um modelo para uma rede neural não linear e não supervisionada que pode ser utilizada para aplicações de classificação e visualização de dados. O mapa de características produzido é uma abordagem de malha de pontos, sendo vetores similares dos dados de entrada agrupados em pontos próximos na malha de saída do método (SHIEH; LIAO, 2012; PAO; MENG, 1998).

2.4.10 *Generative Topographic Mapping*

"Mapeamento topográfico generativo"[6] é similar ao Mapa de características de Kohonen, ou seja, é uma abordagem de malha de pontos, produzindo uma malha na qual vetores similares estão próximos uns dos outros (PAO; MENG, 1998).

2.4.11 *Autoassociative mapping approach*

"Abordagem de mapeamento auto-associativo" [6] faz uso das características de uma rede neural MLP, que deve ser treinada com todos os dados de entrada devendo gerar uma cópia deles na saída. A principal característica da rede é que a sua camada interna de neurônios deve ser restrita quanto ao número de nós. Esta restrição impõe a criação de uma representação com um número de dimensões reduzido dos dados de entrada. (PAO; MENG, 1998).

2.4.12 *Equalized Orthogonal Mapping*

Produz resultados similares à NLVC, otimizando o uso da representação no espaço de saída. As matrizes de covariância dos valores de saídas são designadas para serem diagonais e com elementos de mesmo valor (PAO; MENG, 1998).

[6] Em tradução livre do autor.

2.4.13 *Reduced Neighbors Multidimensional Scaling*

Em seu trabalho Pawliczek e Dzwinel (2013) descrevem uma alteração no método MDS que permite ao algoritmo realizar a redução dimensional utilizando apenas um subconjunto da matriz de distâncias do conjunto original de dados, tornando assim a execução do método mais rápida. Para tanto, os autores assumem que o espaço original Y é fixo, ou seja, não pode ser movido nem rotacionado, e que todos os seus elementos são ligados por um conjunto de ligações rígidas. Assim, o número mínimo de ligações que representa a topologia do sistema é dado pela Equação (39).

$$L_{min} = N \cdot M - N(N+1)/2 \tag{39}$$

A relação entre o número mínimo e máximo de ligações é obtida por:

$$\mu = \frac{L_{min}}{L_{max}} = \frac{2N}{M+1} \cdot \left(1 + \frac{N+1}{2M}\right), \tag{40}$$

considerando que $M > N + 1$ e $L_{max} = M(M-1)/2$, então $M = cN$, resultando na relação:

$$\mu = 1/c. \tag{41}$$

Para os casos em que $M \gg N$, ou seja, para c com valor alto, um fração menor do conjunto de vetores representa a estrutura topológica do espaço original. Consequentemente, apenas um conjunto reduzido $S = \{D_{ij}, z = \#D_{ij}\}$ é necessário para realizar a redução dimensional, sendo $z = \#D_{ij}$ a quantidade de ligações, ou dissimilaridades, selecionadas.

Os resultados apresentados indicaram que a aplicação do conjunto reduzido de vetores proporcionou um ganho de performance da ordem de 8,5 a 10 vezes em relação ao tempo de processamento do método MDS clássico. Contudo, para possibilitar o uso deste algoritmo, em arquiteturas de processamento paralelo como as *Graphical Processing Units* (GPUs), foi necessário alterar o método de cálculo da matriz de dissimilaridades, para que o acesso aos valores fosse otimizado, uma vez que a aplicação do método resultou em uma matriz esparsa.

2.4.14 *Particle-based Multidimensional Scaling*

Os autores Dzwinel e Wcisło (2015) apresentam uma modificação ao algoritmo **NR-MDS!** (**NR-MDS!**), apresentado em Pawliczek e Dzwinel (2013), baseando o método MDS em um sistema de partículas para minimizar $V(\|\Delta - D\|)$, que corresponde à Equação de STRESS (23). As partículas no sistema, que representam os vetores

do conjunto original de dados, interagem entre si por um sistema semi-hormônico de forças, conforme mostra a Equação (42).

$$f_{ij}(X_n) = -grad\left[w_{ij}\left(D_{ij}^k - r_{n,ij}^k\right)^m\right] \tag{42}$$

Cada par de partículas i e j, no instante n estão separadas pela distância $r_{n,ij}^k = \|x_{n,i} - x_{n,j}\|$, e a força atuando nelas corresponde ao somatório das forças pertencentes ao conjunto de seus vizinhos, determinados aleatoriamente. Adicionalmente, é inserido um fator de amortecimento para que o sistema convirja para um estado de energia potencial mínima $E(X) = V(\|\Delta - D\|)$.

Segundo os autores, o uso da Equação de Stress (8) possibilita ao método atingir um valor de mínimo melhor que outros métodos baseados no cálculo do gradiente dos vetores. Os resultados obtidos, ao aplicar o algoritmo em bases de dados com características distintas, mostrou uma melhor separação das classes existentes no conjunto, além de considerável ganho de performance devido a utilização do conjunto mínimo de vetores para os cálculos.

2.4.15 Coordenadas Estrela

O sistema de coordenadas estrela tem como objetivo fornecer um sistema de fácil entendimento para a análise de dados multidimensionais, criando uma projeção 2D do conjunto de dados e fornecendo uma série de ferramentas para que os usuários possam explorar as relações existentes no conjunto de dados de entrada (KANDOGAN, 2000).

Kandogan (2000) descreveu em seu trabalho as coordenadas estrela como um plano no qual os eixos do sistema multidimensional são dispostos de forma circular, compartilhando uma origem comum e separados, um do outro, com o mesmo ângulo. Este sistema pode ser mapeado para o plano de coordenadas cartesianas com a definição de um ponto de origem $O_n(x, y) = (o_x, o_y)$ e uma sequência de vetores $A_n = \langle \vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n \rangle$ que representam os eixos do espaço *n*-dimensional.

Cada um dos pontos D_j do conjunto de entrada é mapeado para o plano cartesiano a partir da soma do vetor unitário de cada eixo, multiplicado pelo valor correspondente a esta coordenada do ponto *n*-dimensional, conforme a Equação (43):

$$P_j(x, y) = \left(o_x + \sum_{i=1}^{n} u_{xi}\dot{(}d_{ji} - min_i),\ o_y + \sum_{i=1}^{n} u_{yi}\dot{(}d_{ji} - min_i)\right), \tag{43}$$

onde, $D_j = (d_1, d_2, \ldots, d_n)$, $|\vec{u}_i| = \dfrac{|\vec{a}_i|}{\max_i - \min_i}$, $\min_i = \min\{d_{ji},\ 0 \leq j < |D|\}$ e $\max_i = \max\{d_{ji},\ 0 \leq j < |D|\}$.

2.4.16 Acurácia dos resultados na redução dimensional

Levando em consideração que cada um dos métodos de RD está melhor adaptado a um determinado padrão nos dados de origem, vários estudos foram realizados com o intuito de verificar o desempenho e acurácia de cada algoritmo, além de métricas de avaliação, ou sistemas de verificação e comparação dos resultados (MARTINS et al., 2014; MOKBEL et al., 2013; van der MAATEN; POSTMA; van den HERIK, 2009; ISENBERG et al., 2013).

O resultado de um método para realizar RD depende de como são consideradas as medidas de similaridade entre os elementos do conjunto. Os métodos lineares, como MDS e PCA utilizam relações globais. Já métodos como SPE e Isomap levam em consideração as distâncias locais, ou seja, consideram as características de relacionamento entre os elementos mais próximos uns dos outros. Estas abordagens geram resultados com distintos graus de preservação das distâncias entre pares de elementos no espaço m-dimensional.

Os autores Martins et al. (2014) apresentam um sistema para avaliar diversos aspectos do processo de RD, no qual as relações entre as características dos conjuntos n e m-dimensionais podem ser estudadas de acordo com as seguintes definições:

Falsos vizinhos Considerando um ponto x_i e sua projeção y_i, para que ocorra a preservação de sua vizinhança, todos os vizinhos de y_i também devem estar próximos de x_i. Caso exista um ponto y_j vizinho a y_i, correspondente a um ponto x_j que não pertence à vizinhança de x_i, então y_j é considerado como um *falso vizinho* de y_i.

Vizinhos desaparecidos Caso exista um ponto x_j vizinho a um ponto x_i, e cuja projeção y_j não pertence à vizinhança de y_i, então y_j é considerado um *vizinho desaparecido* de y_i.

Grupos Os conceitos de *falsos vizinhos* e *vizinhos desaparecidos* podem ser generalizados para grupos, considerando que grupos de elementos próximos no espaço n-dimensional devem formar também, grupos de elementos no espaço m-dimensional. Assim, pode-se generalizar *falsos vizinhos* em *falsos membros* e *vizinhos desaparecidos* em *membros desaparecidos*.

No trabalho de Mokbel et al. (2013), cujo objetivo é aprimorar a avaliação baseada em matrizes de ordem, são apresentadas uma revisão e a taxonomia para critérios de avaliação de métodos de RD, levando em consideração que a formalização das avaliações podem ajudar na comparação entre diferentes métodos de RD ou obter informações qualitativas a respeito de uma determinada visualização obtida a partir de RD.

2.5 COMPUTAÇÃO GENÉRICA EM UNIDADES DE PROCESSAMENTO GRÁFICO

Segundo Kurzweil (1999) a evolução passou a ser uma variante na qual o ser humano é o responsável, chamada tecnologia. Em seu livro, o autor cita um conjunto de diretrizes a respeito da evolução da tecnologia, elaboradas por Arthur C. Clark:

> *- Quando um cientista afirma que algo é possível, ele está quase certamente correto. Quando afirma que algo é impossível, ele muito provavelmente está errado.*
> *- A única maneira de descobrir os limites do possível é se aventurar um pouco além, para o impossível.*
> *- Qualquer tecnologia suficientemente avançada é indistinguível da magia.*[7]
>
> (As três leis da tecnologia de Arthur Clarke) (KURZWEIL, 1999, p. 14)

A partir destas definições em conjunto com um estudo sobre a evolução em nosso planeta e também considerando o desenvolvimento das tecnologias do processamento da informação, Kurzweil (1999) elabora a seguinte lei:

> *The Law of Accelerating Returns*: As order exponetially increases, time exponetially speeds up (that is, the time interval between salient events grows shorter as time passes). (KURZWEIL, 1999, p. 30)

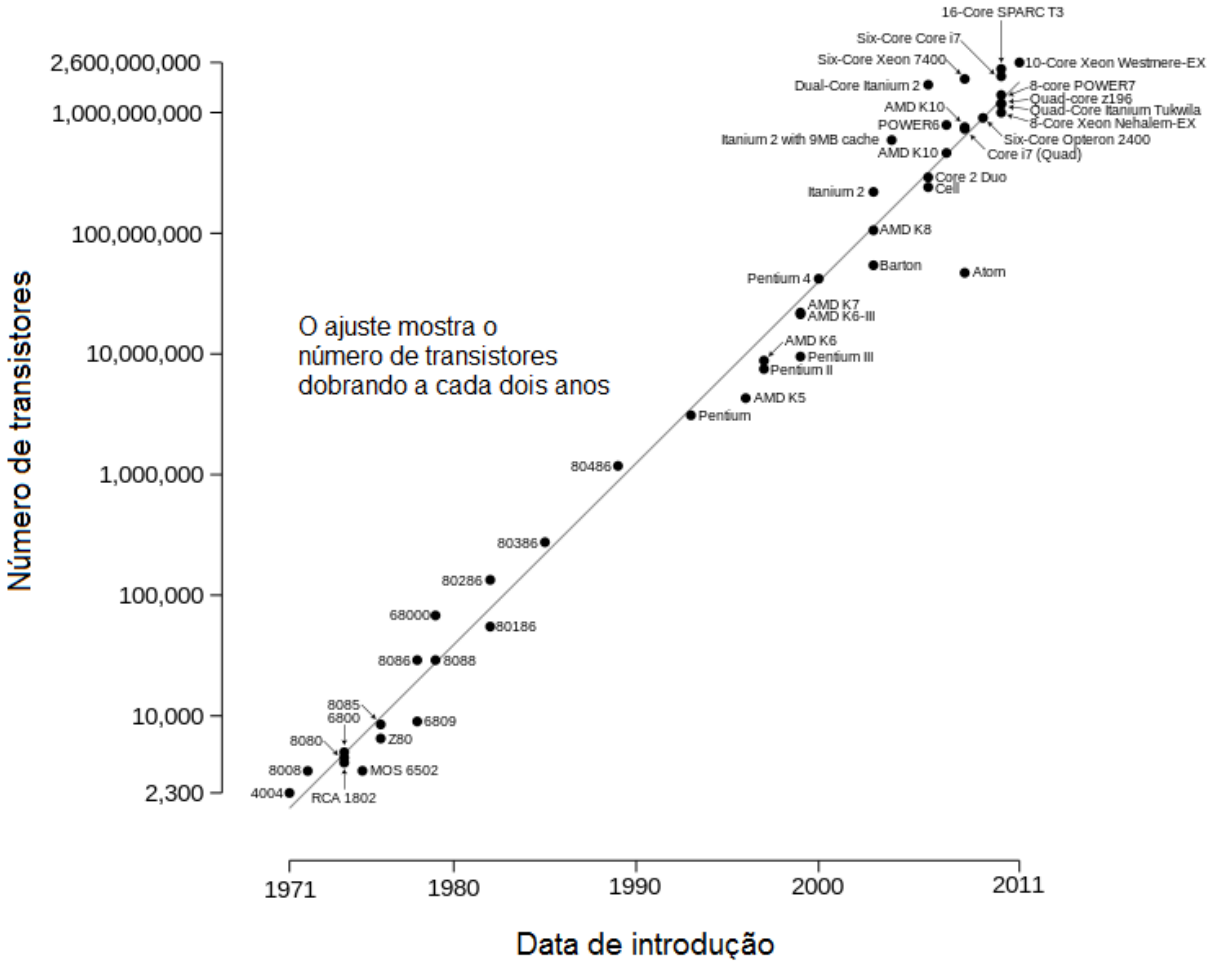
Esta Lei de "Retornos Acelerados" representa o processo de desenvolvimento de uma forma mais geral que a Lei de Moore, descrita adiante, levando em consideração não apenas as técnicas de fabricação de circuitos integrados em silício, mas todo o processo evolutivo da tecnologia.

Atendo-se à evolução dos computadores, a partir do desenvolvimento da arquitetura de Von Neumann, na qual os computadores têm os seguintes componentes principais: memória, unidade lógica e aritmética, unidade de controle e o dispositivo de entrada e saída. A união da unidade lógica e aritmética com a unidade de controle formam o que hoje é conhecido como *Central Processing Unit* (CPU)[8], cuja evolução durante as últimas décadas é governada pela lei de Moore. Esta lei prevê um aumento anual de aproximadamente 60% no número de transistores que podem ser colocados em um chip (TANENBAUM, 2007). A tendência explicitada na lei de Moore pode ser observada no ajuste de curva presente na Figura 10, elaborado a partir das quantidades de transistores de um conjunto de processadores comerciais de um único chip, no período de 40 anos entre 1.971 e 2.011 (WIKIPEDIA, 2015).

Assim, tanto a Lei de Moore quanto a Lei dos Retornos Acelerados corroboram o rápido desenvolvimento dos computadores, mas as demandas impostas a eles

[7] Em tradução livre do autor.
[8] Unidade Central de Processamento (UCP)

FIGURA 10 – Quantidade de transistores segundo a Lei de Moore



FONTE: Adaptado de Wikipedia (2015)

também estão aumentando. E para poder continuar a fornecer maior poder de processamento os engenheiros estão recorrendo cada vez mais a soluções que envolvem o paralelismo, ou seja, máquinas em que mais de uma atividade é executada simultaneamente (TANENBAUM, 2007).

Uma classificação dos tipos de máquinas paralelas corresponde à combinação entre as seguintes situações: a existência de apenas um fluxo de instruções e a existência de um ou mais fluxos de dados, resultando em quatro possíveis situações (TANENBAUM, 2007; PATTERSON; HENNESSY, 2005):

***Single Instruction Single Data* (SISD)** corresponde a um fluxo de instruções e um fluxo de dados, e representa a arquitetura da máquina de Von Neumann.

***Single Instruction Multiple Data* (SIMD)** é uma máquina que possui apenas uma unidade de controle e muitas unidades de execução para poder tratar muitos fluxos de dados simultaneamente.

***Multiple Instruction Single Data* (MISD)** é uma categoria de computadores contro-

versa, e não há consenso sobre a sua existência. Alguns pesquisadores consideram os processadores SISD que implementam *pipeline* pertencentes a esta categoria.

***Multiple Instruction Multiple Data* (MIMD)** caracterizada como máquinas formadas por múltiplas CPUs independentes que operam como partes de um sistema maior.

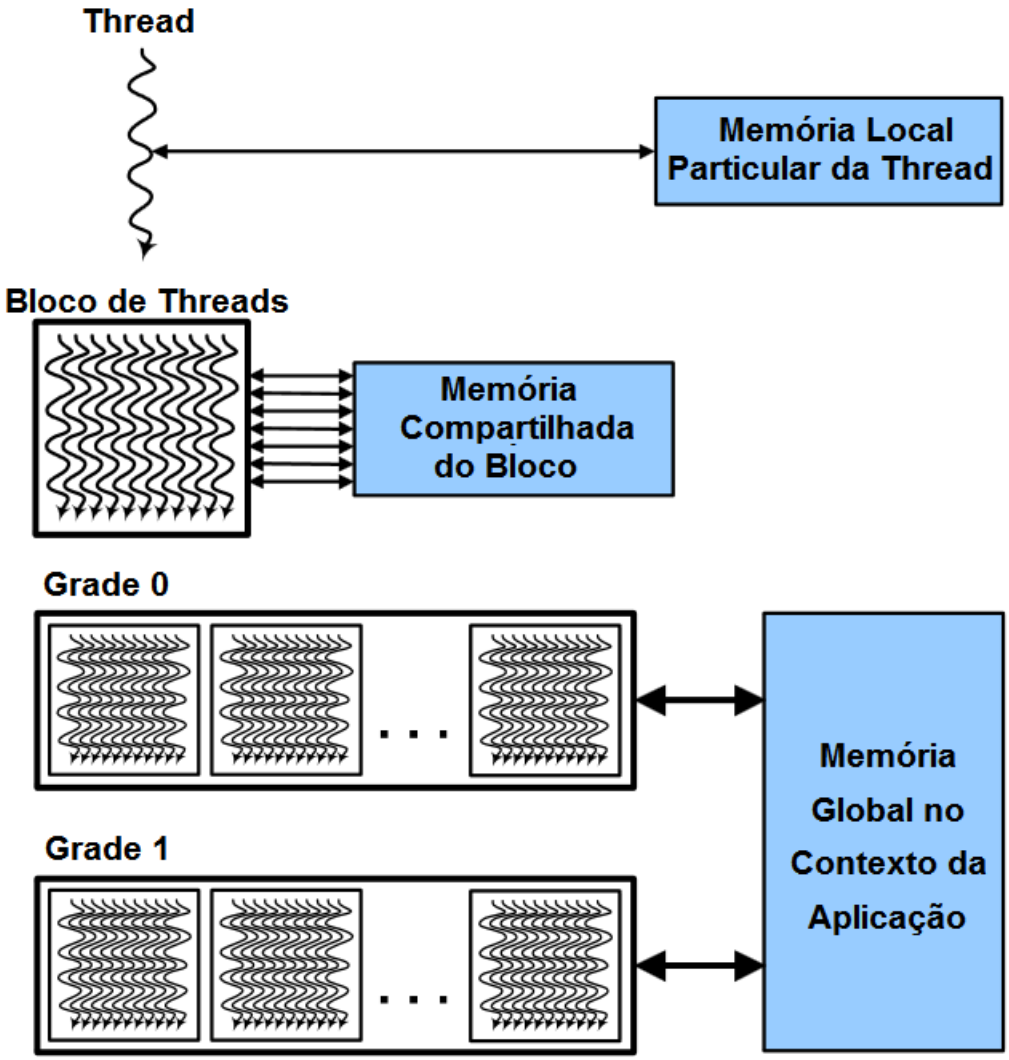
De acordo com o exposto em NVidia (2009), os computadores pessoais disponíveis atualmente, que são equipados com uma CPU, cuja arquitetura provê múltiplos núcleos de processamento, que corresponde a um sistema MIMD, e também utilizando uma unidade de processamento gráfico GPU, correspondente à uma máquina SIMD, disponibilizam um ambiente de computação heterogêneo, pois combinam duas unidades de processamento com arquiteturas distintas em um mesmo computador.

O objetivo inicial do desenvolvimento das unidades gráficas era fornecer aos usuários um equipamento capaz de gerar imagens de alta qualidade em tempo real, mas evoluiu para fornecer um ambiente de processamento de alta performance com operações em ponto flutuante. Estas unidades são classificadas de maneira especial, sendo chamadas de *Single Instruction Multiple Threads* (SIMT), nas quais um conjunto de linhas de execução concorrentes e independentes executam as mesmas instruções sobre um conjunto de dados, que é diferente para cada *thread* em execução (NVIDIA, 2009).

Para aproveitar o desempenho oferecido por estes diferentes dispositivos, cada fabricante disponibiliza um conjunto de ferramentas para a sua programação. No intuito de simplificar o desenvolvimento de aplicações para este ambiente heterogêneo, um consórcio de empresas chamado Khronos Group foi formado, e em 2009 lançou a especificação da primeira linguagem de programação paralela chamada OpenCL, que é aberta e multiplataforma, parte de uma arquitetura computacional cujo objetivo é disponibilizar um ambiente no qual seja possível explorar tanto as características de processamento das CPUs quanto das GPUs disponíveis em um computador. Dentre outras, participam deste consórcio as seguintes empresas: Intel, AMD, Apple Inc, IBM Corporation, NVidia e Samsung Eletronics (GROUP, 2015).

Especificamente para os dispositivos da NVidia, está disponível o *Compute Unified Device Architecture* (CUDA), uma solução de hardware e software que permite a execução de programas escritos em linguagens de programação como C/C++, Fortran, OpenCL entre outras. Uma aplicação CUDA chama *kernels* paralelos, que executam um conjunto de *threads*. As *threads* de um *kernel* são organizadas em blocos, que por sua vez são agrupados em grades, como mostra a Figura 11.

FIGURA 11 – Hierarquia CUDA de *threads*, blocos e grades

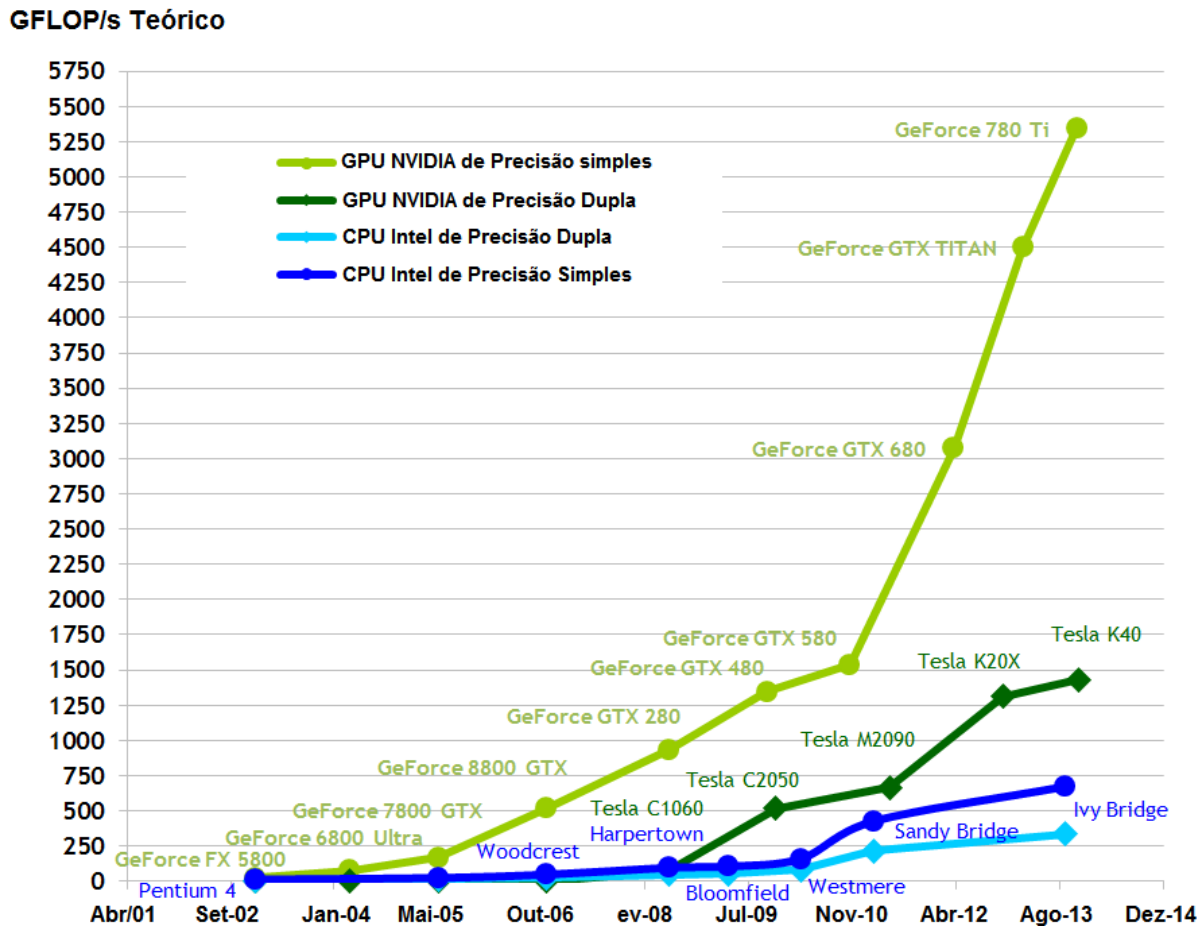


FONTE: Adaptado de NVidia (2009).

Em uma GPU da NVidia, as *threads* CUDA são mapeadas em uma hierarquia de processadores existentes na GPU. Um *Streamming Multiprocessor* (SM) executa um ou mais blocos de *threads*, que são agrupados em conjuntos chamados *warps*, compostos de 32 *threads*. Uma GPU da arquitetura Fermi é composta por 512 núcleos CUDA, que executam uma instrução inteira ou em ponto flutuante por “ciclo de clock”. Estes 512 núcleos estão divididos em 16 SM com 32 núcleos cada um (NVIDIA, 2009).

Aplicações de diversas áreas demandam cada vez mais poder computacional, para atender necessidades de processamento de dados, imagens em tempo real, visualizações tanto para entretenimento como para aplicações científicas. Neste sentido o desempenho dos processadores gráficos supera em muito o dos processadores genéricos, que são utilizados nos computadores, como pode ser observado na Figura 12.

FIGURA 12 – Desempenho de GPUs e CPUs



FONTE: Adaptado de NVidia (2014a).

2.5.1 Paralelização da redução do número de dimensões

Os autores Park, Shin e Hwang (2012) propuseram um método para tratar rapidamente grandes volumes de dados utilizando *General-Purpose Computation on Graphics Processing Units* (GPGPU) quando a quantidade de vetores no conjunto requer mais memória do que a disponível para a GPU.

O algoritmo proposto divide o conjunto de dados em conjuntos menores, de tamanho apropriado à quantidade de memória disponível, um novo conjunto, também de tamanho apropriado é formado a partir de amostras dos conjuntos formados na etapa anterior. Para cada uma das novas matrizes é aplicando o método MDS e em seguida os resultados obtidos são unidos para obter-se uma solução aproximada para a redução dimensional do conjunto de dados de entrada. Os passos para o algoritmo de escala multidimensional rápido estão elencados a seguir:

1. Decompor aleatoriamente a matriz de dissimilaridades $\Delta = [\delta_{ij}]$ de ordem $N \times N$, em p submatrizes $D_1, D_2, \ldots, D_p$.

2. Escolher s amostras aleatórias a partir das submatrizes.

3. Unir os vetores amostrados em uma nova matriz de dissimilaridades M_{align} de ordem $s\dot{p} \times s\dot{p}$.

4. Aplicar o método MDS nas matrizes de dissimilaridade $D_1, D_2, \ldots, D_p$ e M_{align}, denotando os resultados por $dMDS_1, dMDS_2, \ldots, d_MDS_p$ e $mMDS$.

5. Extrair os vetores amostrados no passo 2 das matrizes de resultado, obtendo $subdMDS_1, \ldots, subdMDS_p$, bem como $mMDS_1, \ldots, mMDS_p$.

6. Para cada par $subdMDS_i$ e $mMDS_i$, com $i = 1, 2, \ldots, p$, resolver o utilizando o método dos mínimos quadrados a seguinte equação:

$$arg - min_{A_i} ||A_i \dot{s}ubdMDS_i - mMDS_i||$$

, onde $||\cdot||$ corresponde à norma L^2.

7. Transformar linearmente os vetores D_i utilizando $A_i\, dMDS_i = ndMDS_i$.

8. Combinar $ndMDS_1, ndMDS_2, \ldots, ndMDS_p$ para formar a solução aproximada do método MDS para todo o conjunto de entrada.

De acordo com os autores, o método apresenta uma performance de mais de 100 vezes a velocidade do método MDS implementado utilizando C# ou MATLAB em unidades de processamento sequenciais - CPU, com resultados muito precisos comparados aos obtidos com as implementações clássicas do método.

Neste capítulo foram apresentadas as bases conceituais utilizadas no desenvolvimento deste trabalho, juntamente com uma pesquisa sobre temas e algoritmos correlatos, com um seção a respeito da fundamentação de visualização científica no que concerne à visualização de dados e à geração de imagens para representar volumes, e outra seção a respeito dos conceitos de espaços de cores, utilizados na criação de escalas para os dados e também como fonte de valores para os algoritmos de geração de imagens.

Apresentou-se também uma seção com os fundamentos sobre redução dimensional, na qual a lista de métodos e algoritmos descritos fornecem a fundamentação matemática dos trabalhos já desenvolvidos, as características de cada um dos métodos, aplicações, bem como suas vantagens e desvantagens, com especial atenção para o método MDS e suas diferentes implementações, pois tem como uma de suas principais características o fato de garantir a convergência do processo. No caso das Coordenadas Estrela, a vantagem está em sua simplicidade e velocidade de processamento, por se tratar de uma transformação do sistema de coordenadas utilizado.

Na seção sobre computação genérica em GPUs, foram apresentadas as características relevantes da arquitetura destes sistemas computacionais, nas quais algoritmos, cujos conjuntos de dados são representados como matrizes, podem se beneficiar

para a redução do tempo de processamento necessário, como é o caso de muitos dos algoritmos de RD.

No próximo capítulo serão descritos os métodos e materiais utilizados durante o desenvolvimento do trabalho, incluindo também uma descrição dos conjuntos de testes e do ambiente computacional utilizado.

3 MATERIAIS E MÉTODOS

A Redução Dimensional (RD) também faz parte do conjunto de ferramentas disponíveis para a visualização e análise de dados nos quais o número de dimensões excede a capacidade de compreensão humana ou de representação em um determinado dispositivo. Para exemplificar situações nas quais o uso da RD se faz necessária, neste capítulo são descritas as bases de dados, com detalhamento do conteúdo de algumas das matrizes correspondentes aos volumes de dados a serem utilizados para testar os algoritmos. O capítulo inicia com uma seção para apresentar as ferramentas computacionais utilizadas no desenvolvimento e testes dos algoritmos.

3.1 AMBIENTE COMPUTACIONAL

Para a execução da pesquisa e desenvolvimento da solução proposta, a implementação dos algoritmos bem como os testes de seu funcionamento utilizando *General-Purpose Computation on Graphics Processing Units* (GPGPU) é necessária a utilização de microcomputador equipado com placa gráfica NVidia®, habilitada para trabalhar com a plataforma de desenvolvimento *Compute Unified Device Architecture* (CUDA)®. O computador utilizado é um Notebook Asus modelo G75V, equipado com processador Intel® i7-3630QM, 16GB de memória, placa gráfica NVidia Geforce GTX 670MX equipada com 3GB de memória.

O desenvolvimento de aplicações utilizando CUDA requer a utilização do *CUDA Toolkit*, que consiste em um conjunto de ferramentas e bibliotecas que permite o acesso aos recursos de processamento das placas gráficas da marca NVidia® e seus respectivos processadores no desenvolvimento de algoritmos cuja execução pode ser realizada em paralelo. Faz parte destas ferramentas o compilador nvcc, que é utilizado para compilar o código que deve ser executado pela *Graphical Processing Unit* (GPU) (NVIDIA, 2014b). A versão da *CUDA Toolkit* utilizada é a 6.5, disponível a partir de 01/08/2014.

Juntamente com o compilador do *CUDA Toolkit*, faz-se necessária a utilização de outro compilador, para realizar a compilação do código que deve ser executado pela *Central Processing Unit* (CPU) do microcomputador. O compilador utilizado para a implementação dos elementos a serem executados pela CPU é o Microsoft® C/C++ Compiler e o ambiente de desenvolvimento utilizado é o Microsoft Visual Studio 2012.

Para completar o conjunto de ferramentas de desenvolvimento também é necessária a biblioteca de rotinas *CUDA Basic Linear Algebra Subprograms* (cuBLAS), uma implementação da biblioteca *Basic Linear Algebra Subprograms* (BLAS) utili-

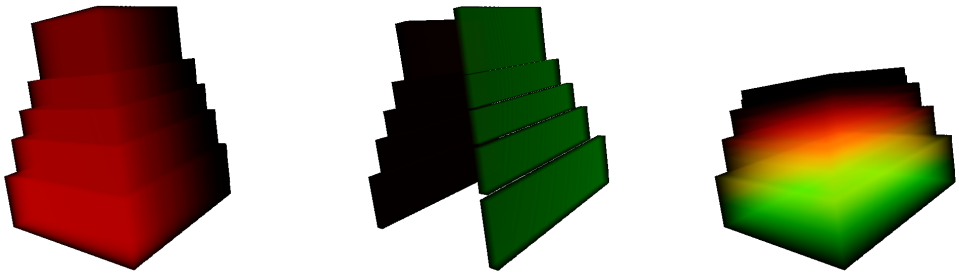
zando o ambiente de execução da CUDA, fornecida juntamente com o *CUDA Toolkit*.

3.2 VOLUME DE TESTE CONTENDO UM BLOCO ARTIFICIAL

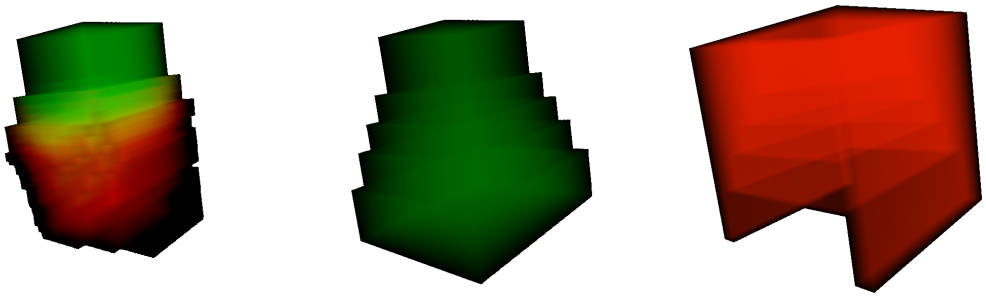
Na Figura 13 é apresentado um bloco tridimensional artificialmente elaborado, constituído por quatro campos escalares, aqui também denominados de dimensões ou variáveis. Os atributos representados não possuem um significado físico específico, mas foram preparados para demonstrar a interação entre dimensões com interpretação, escalas e taxas de variação distintas umas das outras. No restante do texto este volume de dados será denominado Bloco 22.

FIGURA 13 – Quatro campos escalares correspondentes ao volume artificial Bloco 22

(a) Dimensão 1 (b) Dimensão 2 (c) Dimensão 3



(d) Dimensão 4 (e) Volume do Bloco (f) Volume externo ao Bloco

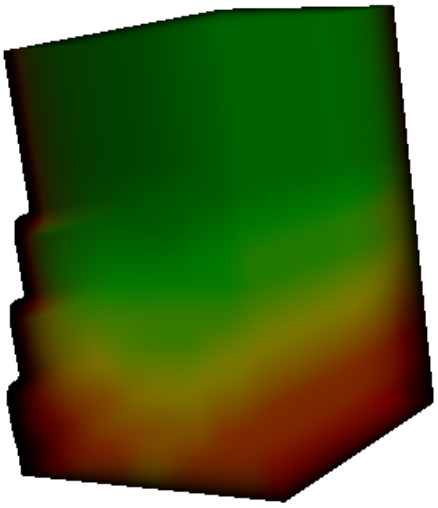


FONTE: Autoria própria

A matriz do volume de dados do Bloco 22 representado na Figura 13, possui dimensões $[22; 22; 22]$ com um total de 982 vetores distintos, ou seja, células da matriz cujos valores aparecem em mais de uma posição são considerados uma única vez.

As dimensões, ou variáveis, representadas nas Figuras 13a e 13c apresentam uma variação linear, cujas direções coincidem cada uma com um dos eixos do sistema de coordenadas cartesiano. Já na Figura 13b a representação demonstra que existem dados apenas em um conjunto de faces do Bloco 22, e na Figura 13d está representada uma variável que apresenta taxa de variação não linear de forma radial, para a qual o centro encontra-se no topo do bloco, em concordância com o conteúdo da Figura 14.

FIGURA 14 – Recorte evidenciando taxa de variação não linear



FONTE: Autoria própria

A Figura 13e apresenta o que pode ser considerado como a parte sólida do volume, ou seja, a região do modelo para a qual serão geradas informações, ou representações visuais na imagem de saída. Na Figura 13f está representado o que diz respeito a valores que se encontram na região externa ao sólido, ou em outras palavras, a região do volume para a qual não existem valores a serem apresentados na imagem de saída. Estas imagens servem para demostrar que os volumes a serem processados pelos algoritmos implementados neste trabalho, devem ser representados por matrizes que compreendam uma região do espaço Euclidiano limitada por seis planos. Consequentemente, caso o volume de dados a ser processado e representado não ocupe completamente um volume do espaço Euclidiano com estas características, vão existir regiões que podem ser consideradas como vazias, sendo a Figura 13f um exemplo disto.

De forma similar ao apresentado para o Bloco 22, foram criados outros vinte e cinco (25) volumes de dados artificiais, com tamanhos variando de [5; 5; 5] a [128; 128; 128] e cujas dimensões estão entre 2 e 6. Para exemplificar diferentes características abordadas nos volumes de testes na Figura 15 são apresentados três diferentes configurações para os volumes de dados artificiais.

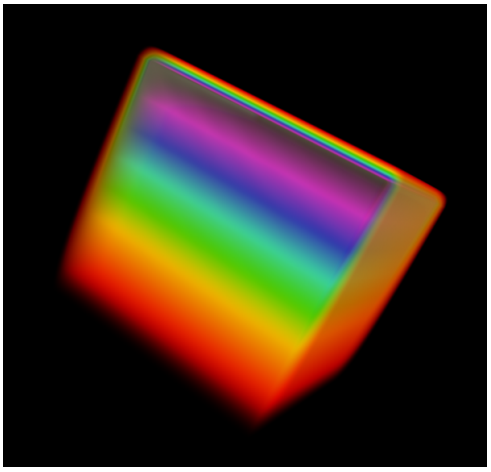
Para facilitar a compreensão dos resultados, as demostrações e explicações descritas no restante do trabalho serão baseadas no volume definido como Bloco 22.

3.3 BASE DE DADOS RESULTANTE DA SIMULAÇÃO DO FURACÃO ISABEL

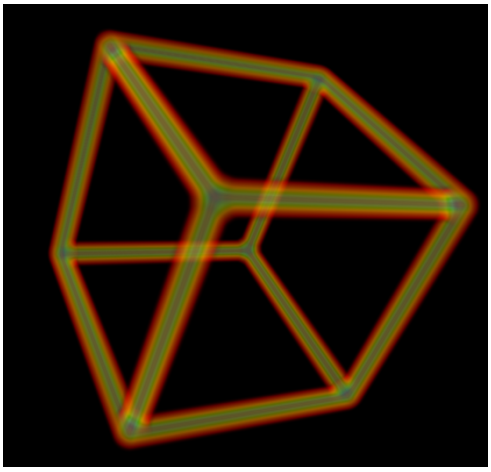
Para a validação do processo de visualização além dos dados representando o bloco da Figura 13, utilizou-se a base de dados resultante da simulação do Furacão

FIGURA 15 – Exemplos de volumes de dados

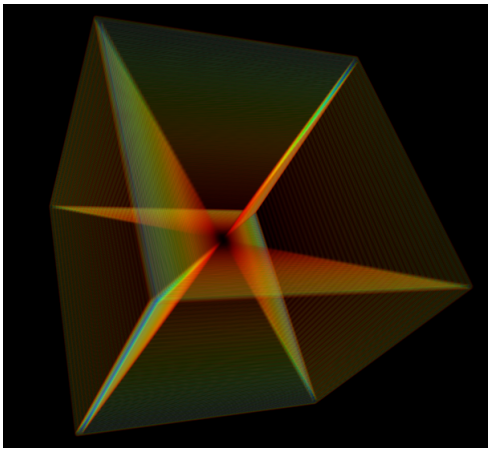
(a) Bloco com variação linear



(b) Conjunto de arestas



(c) Arestas de um conjunto de cubos



FONTE: Autoria própria

Isabel[1], disponibilizada pelo *U.S. National Center for Atmospheric Research* (NCAR). Os dados são compostos por 48 intervalos de tempo, cada um correspondendo a uma matriz de dimensões $[500; 500; 100]$, sendo a primeira dimensão (X) correspondendo a Longitude, no intervalo aproximado de 83°W até 62°W. A segunda dimensão (Y) corresponde a Latitude, com intervalo aproximado de 23,7°N até 41,7°N e a terceira dimensão (Z) corresponde à elevação, ou altura a partir do nível do mar, com intervalo de 0,035 Km até 19,835 Km, contando com 100 níveis igualmente espaçados com intervalos de 0,2 Km (NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR), 2009).

Em cada um dos intervalos de tempo estão disponíveis os valores para 13 variáveis, conforme apresentado na Tabela 1.

As variáveis da base de dados, no trigésimo intervalo de tempo da simulação, podem ser visualizadas na Figura 16. A respeito do conteúdo destas imagens é

[1] http://www.vets.ucar.edu/vg/isabeldata

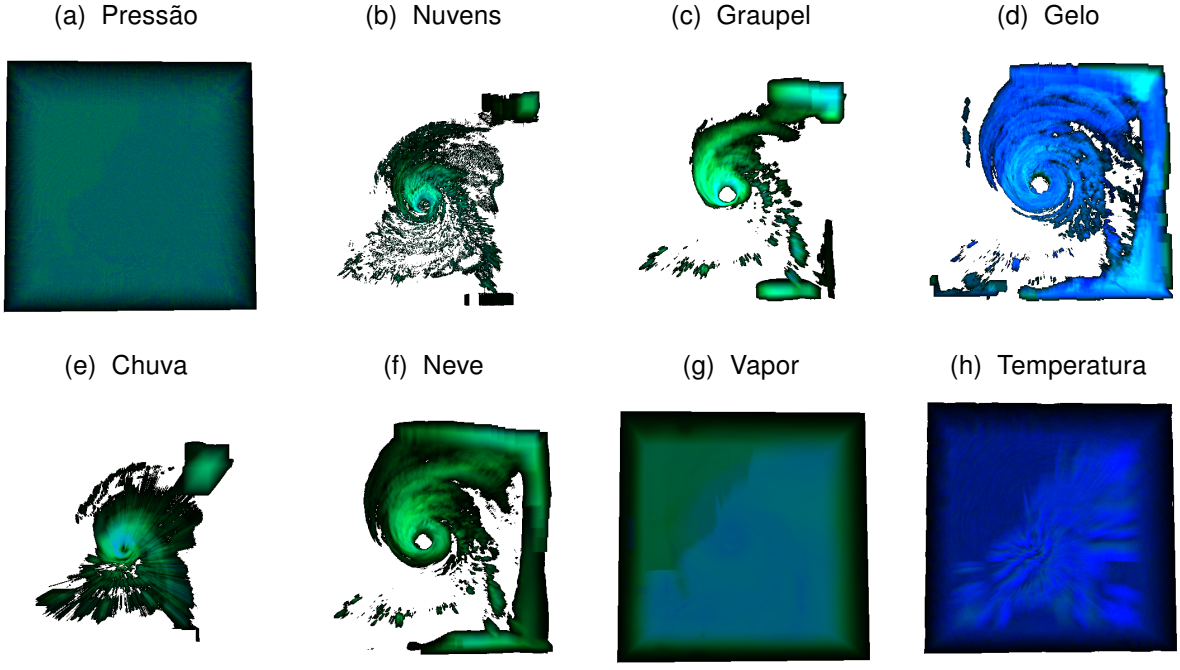
TABELA 1 – Variáveis presentes na base de dados da simulação do Furacão Isabel

Variável	**Descrição**	**Mínimo**	**Máximo**	**Unidade**
QCLOUD	Água em nuvem	0,00000	0,00332	
QGRAUP	Graupel	0,00000	0,01638	kg
QICE	Gelo em nuvem	0,00000	0,00099	kg
QRAIN	Chuva	0,00000	0,01132	kg
QSNOW	Neve	0,00000	0,00135	kg
QVAPOR	Vapor de água	0,00000	0,02368	kg
CLOUD	Peso total da nuvem	0,00000	0,00332	kg
PRECIP	Peso total da precipitação	0,00000	0,01672	kg
P	Pressão (peso da atmosfera)	-5471,85791	3225,42578	Pascal
TC	Temperatura	-83,00402	31,51576	Graus Celsius
U	Componente de vento em X	-79,47297	85,17703	m/s
V	Componente de vento em Y	-76,03391	82,95293	m/s
W	Componente de vento em Z	-9,06026	28,61434	m/s

FONTE: Adaptado de National Center for Atmospheric Research (NCAR) (2009)

interessante verificar que nas componentes de Pressão (P), Vapor (QVAPOR) e Temperatura (T) é possível visualizar o contorno da costa leste da América do Norte, nas demais componentes o contorno não aparece em decorrência do fato de que os dados de cada uma não apresentam valores para as baixas altitudes, conforme o conteúdo da Figura 17.

FIGURA 16 – Projeção individual das variáveis, ou dimensões, do conjunto de dados do Furacão Isabel no tempo 30, com uma vista superior

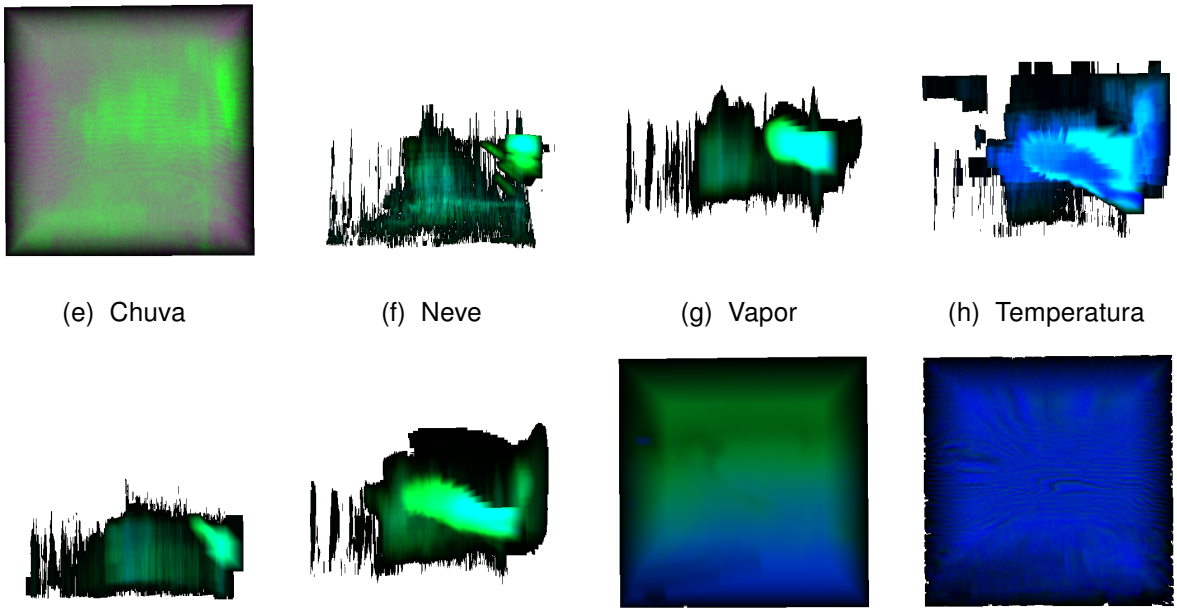


FONTE: Autoria própria

A base de dados da simulação do Furacão Isabel está representada no texto

FIGURA 17 – Projeção individual das variáveis, ou dimensões, do conjunto de dados do Furacão Isabel no tempo 30, com uma vista lateral, no sentido sul-norte

(a) Pressão (b) Nuvens (c) Graupel (d) Gelo



(e) Chuva (f) Neve (g) Vapor (h) Temperatura

FONTE: Autoria própria

pelos dados do trigésimo intervalo de tempo. Porém, testes com os algoritmos foram elaborados a partir dos demais 47 instantes disponíveis na base de dados fornecida pelo NCAR. Além da utilização de outros intervalos de tempo, também foram realizados testes utilizando outras configurações de variáveis, ou dimensões, da simulação.

Para facilitar a compreensão dos resultados, as demostrações e explicações descritas no restante do trabalho serão baseadas no volume definido pelo intervalo de tempo 30 da simulação do furacão.

3.3.1 Recorte dos dados do Furacão Isabel

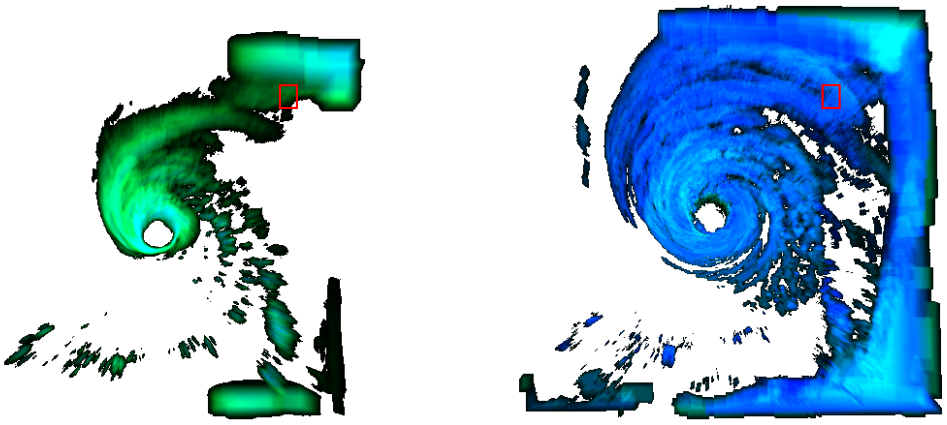
De acordo com o evidenciado na Seção 2.4.2, o algoritmo *Scaling by Mojorizing a Complicated Function* (SMACOF) impõe uma restrição à quantidade de elementos que podem ser processados simultaneamente em relação ao tamanho da memória disponível. Portanto, as matrizes de dados obtidas a partir da base de dados correspondente ao resultado das simulações do Furação Isabel não podem ser processadas utilizando o SMACOF devido ao seu tamanho.

Com o objetivo de contornar esta restrição com relação à quantidade de memória necessária, foram considerados subconjuntos obtidos a partir das matrizes de dados da simulação. A seguir um destes subconjuntos é descrito, evidenciando principalmente os tamanhos necessários em relação às características da implementação e do equipamento utilizado nos testes.

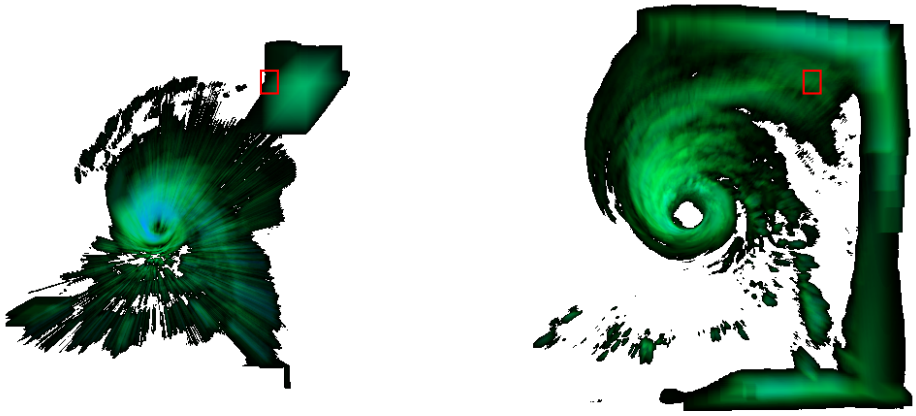
Para que o tamanho da matriz de entrada, formada a partir dos dados no tempo 30 da simulação, seja compatível com a quantidade de memória disponível na placa gráfica, a ser utilizada para a execução o algoritmo SMACOF, foi utilizado um recorte, cuja matriz é de ordem $[27; 20; 35]$ abrangendo a região identificada na Figura 18. São considerados para o teste em questão os valores correspondentes às quantidades de Graupel, Gelo, Chuva e Neve.

FIGURA 18 – Localização do recorte a ser aplicado no conjunto de dados do Furacão Isabel

(a) Graupel (b) Gelo



(c) Chuva (d) Neve



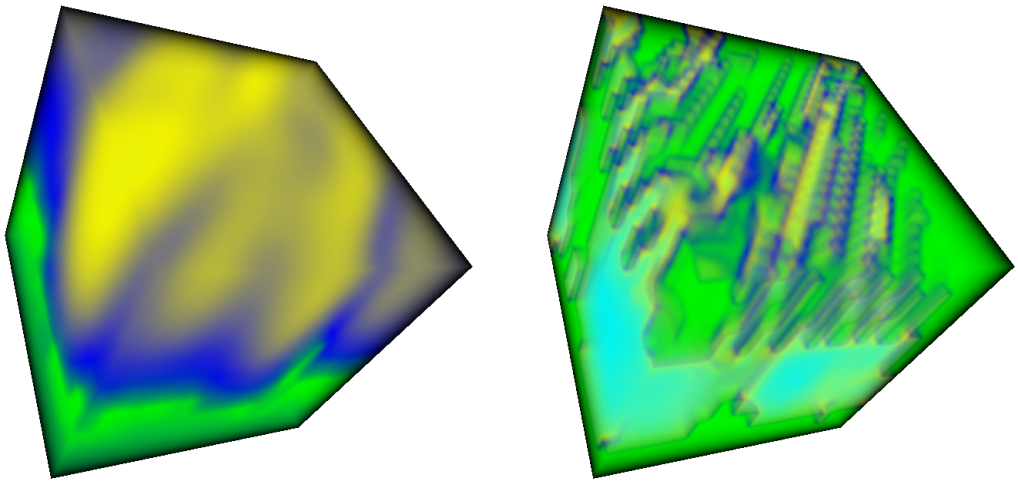
FONTE: Autoria própria

A região delimitada para o teste está identificada nas Figuras 19a, 19b, 19c e 19d. Esta submatriz possui um total de 18.900 elementos dos quais 13.198 não se repetem e, portanto, são utilizados para a composição da matriz de entrada para o algoritmo SMACOF.

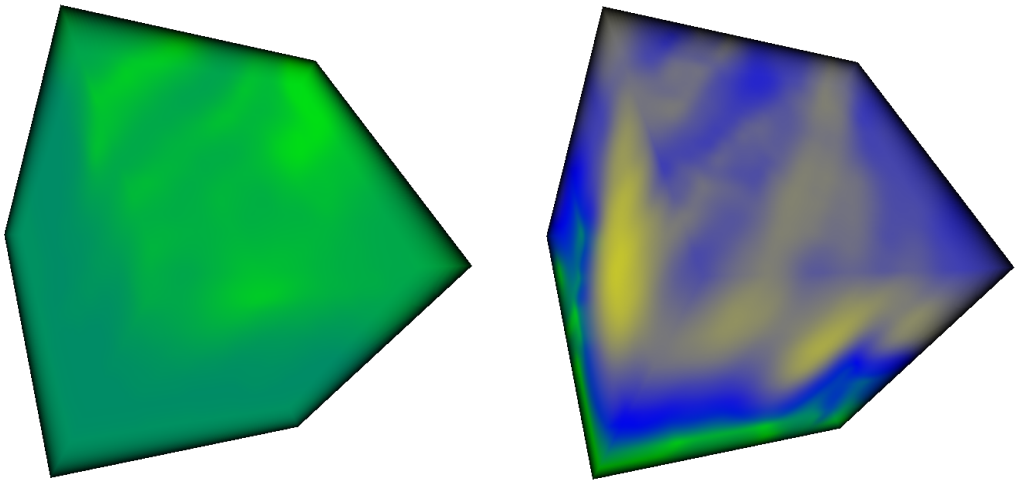
Os volumes correspondentes a cada uma das variáveis apresentadas na Figura 18, após o recorte da matriz, estão reproduzidas na Figura 19.

FIGURA 19 – Volume de dados para teste após recorte do conjunto de dados do Furacão Isabel no tempo 30

(a) Graupel (b) Gelo



(c) Chuva (d) Neve



FONTE: Autoria própria

O recorte da base de dados da simulação do Furacão Isabel que está apresentado nas Figuras 18 e 19 foi utilizado no restante do texto para exemplificar a execução dos algoritmos. Porém, além desta região, várias outras foram consideradas ao testar e executar os algoritmos, sendo utilizados outros intervalos de tempo, outras regiões e, também, distintas configurações de variáveis da simulação.

Neste capítulo foram abordados e descritos os equipamentos e sistemas utilizados na execução do trabalho, juntamente com as bases de dados correspondentes aos volumes empregados durante o desenvolvimento e testes dos algoritmos propostos para realizar a visualização das informações n-dimensionais de forma interativa, conforme os objetivos estabelecidos.

No próximo capítulo serão abordadas e detalhadas quatro abordagens de processamento para realizar a visualização do dados apresentados.

4 VISUALIZAÇÃO DE VOLUMES MULTIDIMENSIONAIS

Neste capítulo são descritos os procedimentos utilizados para o desenvolvimento do protótipo bem como os algoritmos necessários para validar a hipótese apresentada como objetivo do trabalho.

Em primeira instância o problema consiste em criar um método capaz de representar as diferentes classes ou grupos de elementos existentes nos dados a partir da sua projeção em um espaço com um número de dimensões reduzido em comparação ao conjunto original de dados. De acordo com o conteúdo da Seção 2.4, a redução do número de dimensões pode ser realizada por diferentes métodos, com resultados que podem ser considerados em processos tanto de análise quanto de visualização dos dados.

A respeito das técnicas de visualização de grandes volumes de dados, o processo envolve a análise do conjunto, a determinação dos dados de maior relevância, para em seguida realizar-se uma diminuição do espaço, ou seja, diminuir o número de dimensões do conjunto de dados empregando uma de diversas técnicas com o intuito de agrupar ou simplesmente reduzir o número de dimensões dos dados.

A grande maioria das soluções estudadas não preserva as relações espaciais existentes entre as amostras do conjunto de entrada. Uma exceção é o trabalho de Lawrence et al. (2011), no qual estas propriedades são mantidas devido ao fato de os dados de entrada serem imagens multibanda, e a saída deve apresentar o conteúdo com as mesmas distribuições espaciais existentes nas imagens de entrada. Esta mesma concepção poderá ser aplicada a um conjunto de dados, que não seja relacionado a informações do espectro eletromagnético, como no caso dos dados de condições estruturais de uma estrutura de concreto ou os valores correspondentes às condições meteorológicas de um fenômeno climático como um furacão, por exemplo. A proposta apresentada pelos autores resulta em uma transformação do espaço *n*-dimensional para um espaço *m*-dimensional, com $m \ll n$, como definido na Seção 2.4. Nos algoritmos apresentados a seguir são considerados os casos em que $m = 1$, conforme a (44) e $m = 2$ de acordo com a (45).

$$T : \mathbb{R}^n \rightarrow \quad \mathbb{R} \tag{44}$$

$$T : \mathbb{R}^n \rightarrow \quad \mathbb{R}^2 \tag{45}$$

Uma segunda definição adotada neste trabalho é que o espaço para o qual se está projetando o volume de dados original é um espaço de cores e segue as definições apresentadas na Seção 2.2. Para o caso em que o espaço de destino possui

apenas uma dimensão assume-se que os valores pertencem a uma escala de cores como a exibida na Figura 20.

FIGURA 20 – Escala de cores utilizada para mapear os elementos da projeção resultante da RD

FONTE: Autoria própria

De acordo com as características do processo de redução dimensional, o espaço destino, ou seja, o espaço com número reduzido de dimensões, é contínuo e ilimitado, ao contrário dos espaços de cores utilizados computacionalmente, que são limitados e discretos. Para que seja possível então, que a redução de dimensões leve a um espaço de cores, os seus resultados devem ser normalizados e discretizados de tal modo que seja possível interpretar os dados como coordenadas no sistema de cores.

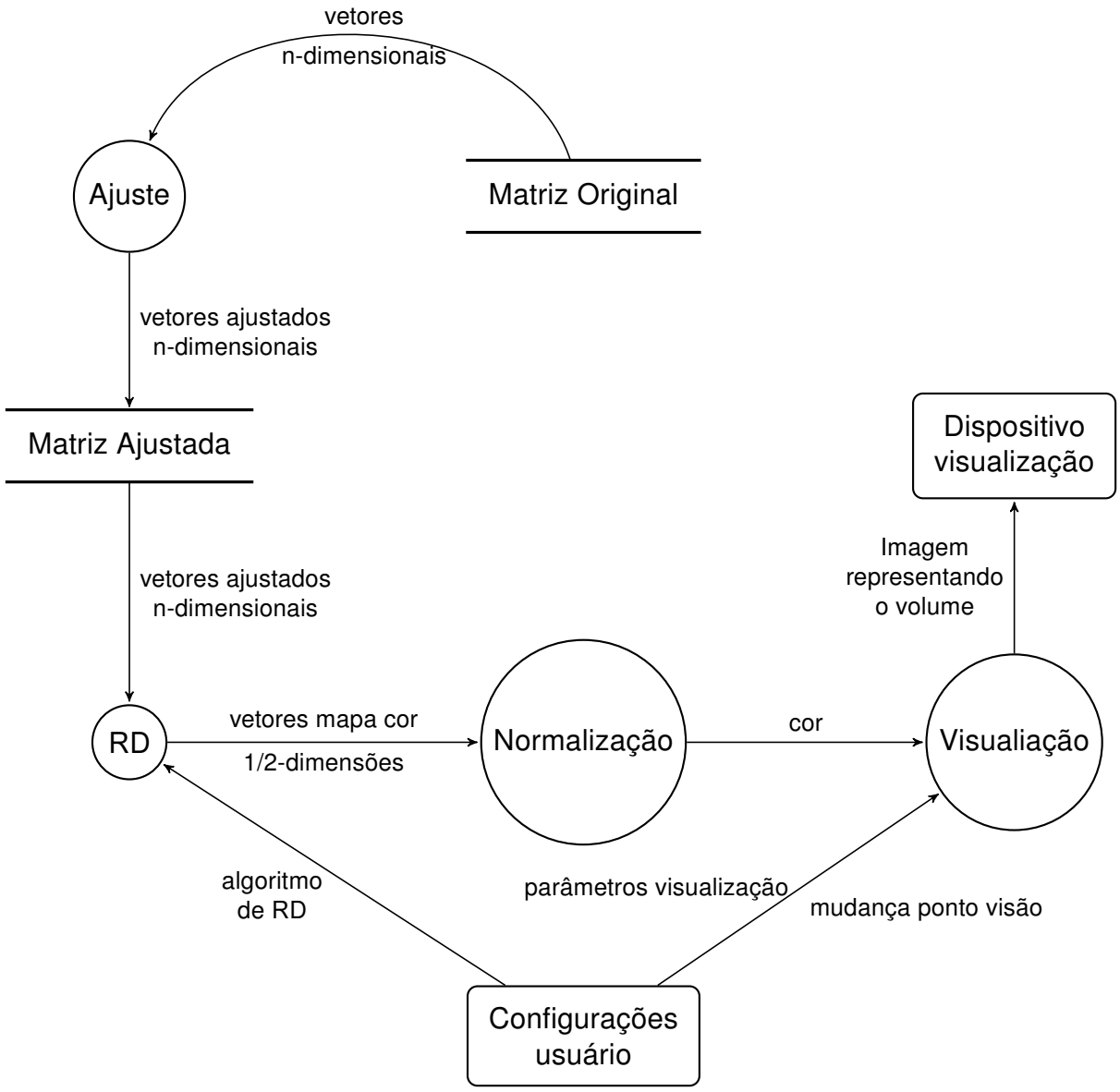
A determinação de qual espaço de cores será utilizado como alvo da redução dimensional afeta também a etapa seguinte, que é a visualização do volume de dados, pois de acordo com o tipo de representação de cores utilizada, o cálculo da integral de visualização de volumes, apresentada na Seção 2.1.1, deve ser adequado para levar em consideração o valor das cores no processo iterativo que determina o valor final das intensidades luminosas.

Neste processo deve ser levada em consideração a capacidade de mesclar corretamente as matizes de diferentes cores, de modo a permitir a representação gráfica de situações em que o angulo de observação, ou a disposição dos dados no espaço original levem a diferentes características estarem presentes no caminho de um mesmo raio de luz a ser calculado, e portanto uma nova cor deve ser utilizada, ou gerada para representar esta região na imagem de saída.

Em conjuntos de dados multidimensionais é comum existir diferenças de unidades ou de escala entre os valores correspondentes a cada uma das dimensões, e para permitir que diferentes grandezas sejam visualizadas de forma simultânea, elas devem passar por um processo de normalização, com o propósito de que dimensões, cujos valores apresentem diferenças de uma ou mais ordens de grandeza, não tornem os demais dados, presentes no volume, irrelevantes.

As etapas do processo proposto, para visualização de volumes de dados n-dimensionais, estão representadas no Diagrama de Fluxo de Dados (DFD) da Figura 21. Os algoritmos implementados em cada um dos processos que fazem parte da solução proposta serão discutidos separadamente nas próximas seções, de acordo com o técnica de Redução Dimensional (RD) a ser empregada.

FIGURA 21 – Diagrama de Fluxo de Dados para a solução proposta



FONTE: Autoria própria

No fluxo de dados, o processo *Ajuste* é responsável por adequar os valores e distribuição espacial dos vetores de dados da matriz original. O processo de *RD* utiliza um dos algoritmos de acordo com a escolha do usuário para reduzir o conjunto de dimensões/variáveis dos dados para 1 ou 2 dimensões. Já a *Normalização* é responsável por traduzir os resultados da RD em um conjunto de valores de cores, representando cada uma das regiões do volume para a qual existem informações na *Matriz Original*. Por último, o processo *Visualização* utiliza o mapa de cores resultante para criar a *Imagem* que representa o volume n-dimensional, de acordo com o ponto de vista indicado pelo usuário.

4.1 PROJEÇÕES UNIDIMENSIONAIS

O procedimento, apresentado em Matrakas e Scheer (2016a), para realizar a projeção de um espaço *n*-dimensional em um espaço unidimensional pode ser dividido em quatro etapas, elencadas a seguir:

1. Preparar os dados para a redução dimensional, ajustando as amostras das diferentes dimensões para ocuparem a mesma grade;
2. Aplicar o processo de redução dimensional, para obter uma projeção do volume em espaço reduzido;
3. Normalizar o resultado da redução dimensional, de maneira que o novo volume de dados possa ser interpretado como coordenadas em um espaço de cores;
4. Gerar a representação gráfica, ou imagem, que representa o volume de dados, ou seja, realizar a visualização do volume.

Na primeira etapa, os dados do volume original devem ser ajustados de maneira que todas as dimensões, ou variáveis, compartilhem o mesmo domínio. Além deste ajuste nas amplitudes dos valores é necessário também que as amostras para cada uma das dimensões estejam em conformidade, isto é, se a distribuição dos sensores de aquisição dos dados para cada uma das dimensões possuir diferentes resoluções, as grades, ou seja, o espaço amostral, de todas devem ser ajustados de maneira que apenas uma grade de dados represente o volume nos passos seguintes.

A segunda etapa consiste na execução do algoritmo de redução dimensional, com o objetivo de projetar os dados do volume, ajustado na primeira fase, em um espaço unidimensional, que no terceiro passo será normalizado de modo a ser interpretado como um espaço de cores. Cada um dos vetores do volume de dados original é então substituído pelo código de cor resultante da normalização, e uma imagem do volume pode ser então criada a partir destas informações.

Finalizando o processo, o quarto passo corresponde à geração da imagem para representar o conteúdo da matriz original, levando em consideração os valores de todas as dimensões, ou variáveis, que compõem o conjunto de dados. O algoritmo de visualização não leva em consideração a dispersão nem a emissão de luz pelo meio a ser representado. Portanto, a imagem é gerada a partir das características de absorção da luz em cada segmento do volume de dados.

Uma característica importante a ser ressaltada no processo proposto diz respeito à informação espacial dos dados, ou seja, à posição no espaço onde determinado valor ocorre. Apesar de serem dimensões do conjunto de dados, as coordenadas

não compõem a matriz de dados a ser apresentada ao método de redução dimensional, isto porque cada valor do campo escalar ocupa a posição correspondente à sua localização na grade de dados, de forma similar ao contexto do trabalho apresentado por Lawrence et al. (2011).

Para exemplificar esta característica a Equação (46) define um volume de dados com dimensões $[3; 3; 3]$ formado por dois campos escalares. Esta mesma matriz é apresentada na Figura 22 de forma a evidenciar o posicionamento dos elementos no volume formado pelos dados. Na representação da Equação (46), d corresponde ao índice dos campos escalares, ou seja, às colunas de matrizes e na Figura 22 corresponde ao índice dos valores em cada uma das células. z corresponde ao posicionamento vertical tanto na equação quanto na figura, x é o índice das linhas das matrizes na representação da Equação (46) e na figura corresponde ao posicionamento horizontal, já y indica as colunas das matrizes na equação e a profundidade na figura.

$$M(d,z,x,y) = \left[\left[\left[\begin{bmatrix} 235 & 175 & 235 \\ 235 & 000 & 235 \\ 235 & 175 & 235 \end{bmatrix} \begin{bmatrix} 175 & 000 & 175 \\ 000 & 000 & 000 \\ 175 & 000 & 175 \end{bmatrix} \begin{bmatrix} 235 & 175 & 235 \\ 235 & 000 & 235 \\ 235 & 175 & 235 \end{bmatrix} \right] \left[\begin{bmatrix} 100 & 110 & 120 \\ 170 & 000 & 130 \\ 160 & 150 & 140 \end{bmatrix} \begin{bmatrix} 090 & 000 & 110 \\ 000 & 000 & 000 \\ 150 & 000 & 130 \end{bmatrix} \begin{bmatrix} 080 & 090 & 100 \\ 150 & 000 & 110 \\ 140 & 130 & 120 \end{bmatrix} \right] \right] \right] \quad (46)$$

O procedimento adotado para projetar um volume em um espaço unidimensional é formado pelos seguintes passos:

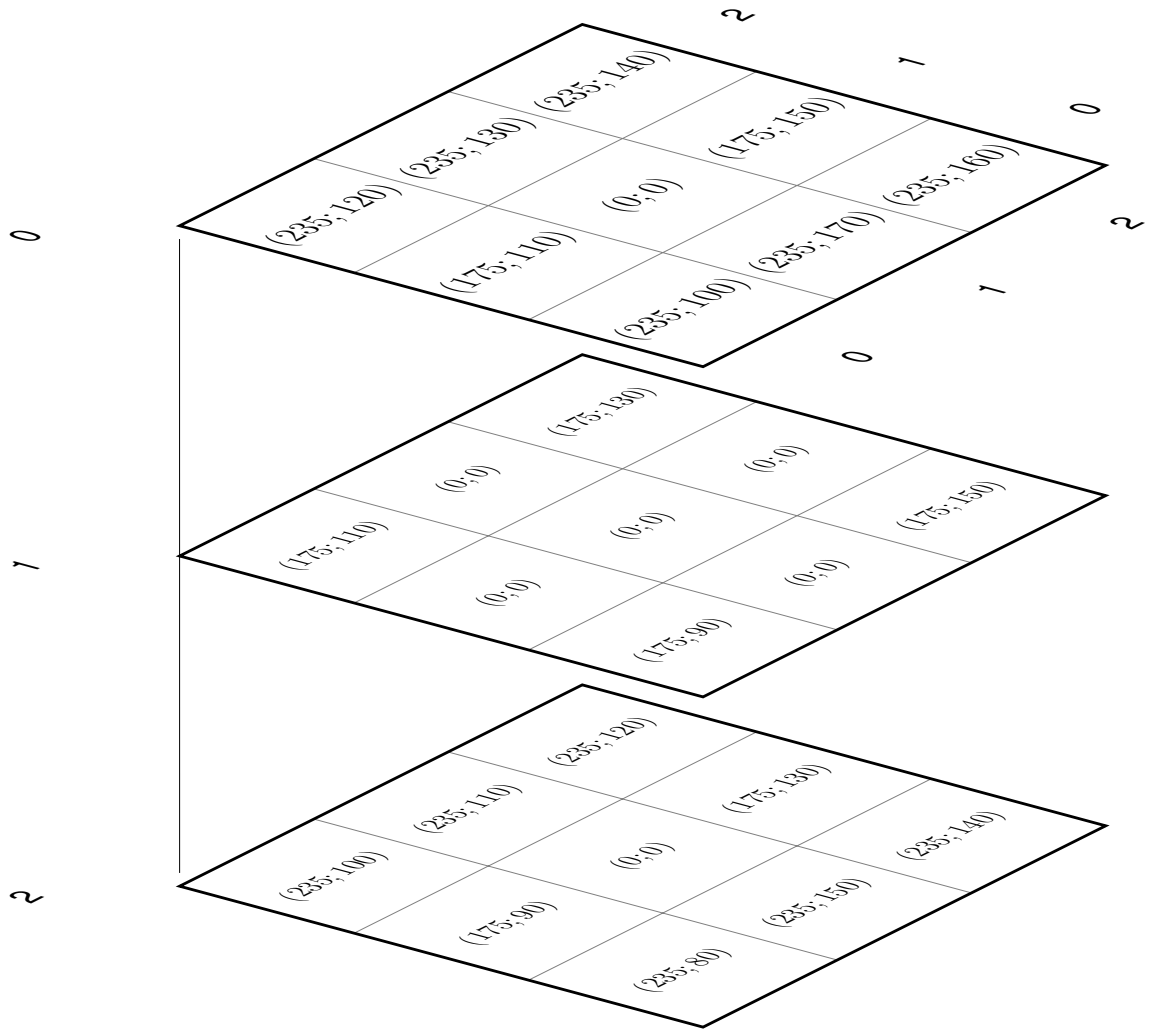
1. Normalizar os dados, individualmente para cada uma das dimensões, utilizando o intervalo de 0 a 255 (com o objetivo de representar o volume de dados de maneira otimizada);

2. Verificar a existência de mais de uma ocorrência para cada um dos elementos, registrando a posição em que eles ocorrem na matriz de entrada em uma matriz de mapeamento e formando uma nova matriz de dados com elementos singulares;

3. Calcular a matriz de distâncias entre os elementos distintos da matriz de elementos singulares;

4. Realizar a projeção da matriz de elementos singulares para o espaço de destino m, unidimensional;

FIGURA 22 – Representação 3D de um volume de dados com um vetor bidimensional em cada posição



FONTE: Autoria própria

5. Aplicar o algoritmo *Scaling by Mojorizing a Complicated Function* (SMACOF) utilizando a matriz de distâncias do passo 3 e a projeção encontrada no passo 3 como parâmetros de entrada;

6. Considerar os resultados do SMACOF e utilizar os valores nas posições correspondentes da matriz de mapeamento do passo 2;

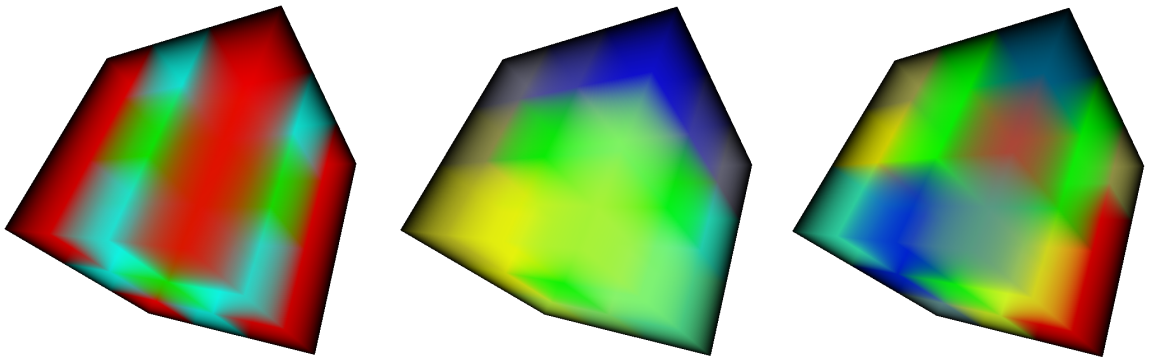
7. Utilizar a matriz de dados reduzida, obtida no passo 6 como entrada para o algoritmo de visualização, para a geração da representação do volume de dados com suas dimensões reduzidas.

O resultado deste algoritmo aplicado na matriz de dados da Figura 22 é apresentado na Figura 23. A Figura 23a corresponde à primeira dimensão da matriz original e a Figura 23b apresenta o conteúdo da segunda dimensão. Na Figura 23c tem-se

o resultado do algoritmo proposto, apresentando as características combinadas dos dados presentes nas duas dimensões de entrada.

FIGURA 23 – Renderização das dimensões da matriz apresentada na Figura 22

(a) Dimensão 1 (b) Dimensão 2 (c) Dimensão MDS



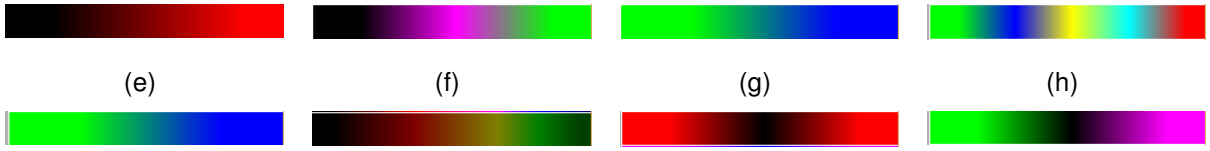
FONTE: Autoria própria

Os valores resultantes são interpretados como índices para uma das escalas de cores representadas na Figura 24, cujos valores variam no intervalo: $[0; 255]$. Nesta etapa, as escalas de cores utilizadas são definidas no espaço de cores RGB, e o processo de geração das imagens a partir do resultado da redução dimensional faz uma soma ponderada dos valores de cada posição do volume, de acordo com as características de opacidade escolhidas de forma interativa.

FIGURA 24 – Escalas de cores utilizadas nas projeções unidimensionais

(a) (b) (c) (d)

(e) (f) (g) (h)



FONTE: Autoria própria

4.2 PROJEÇÕES BIDIMENSIONAIS

As mesmas definições apresentadas na seção anterior, e também em Matrakas e Scheer (2016b), são válidas ao se considerar a projeção de um volume *n*-dimensional em um espaço bidimensional.

O objetivo da redução dimensional com $m = 2$ é melhorar a preservação das relações existentes entre os vetores dos dados originais e buscar manter a integridade dos grupos que ocorrem no volume de dados original.

A RD de um volume para o espaço bidimensional exige que o espaço de origem seja pelo menos de dimensão 3, para que haja uma redução do número de dimensões.

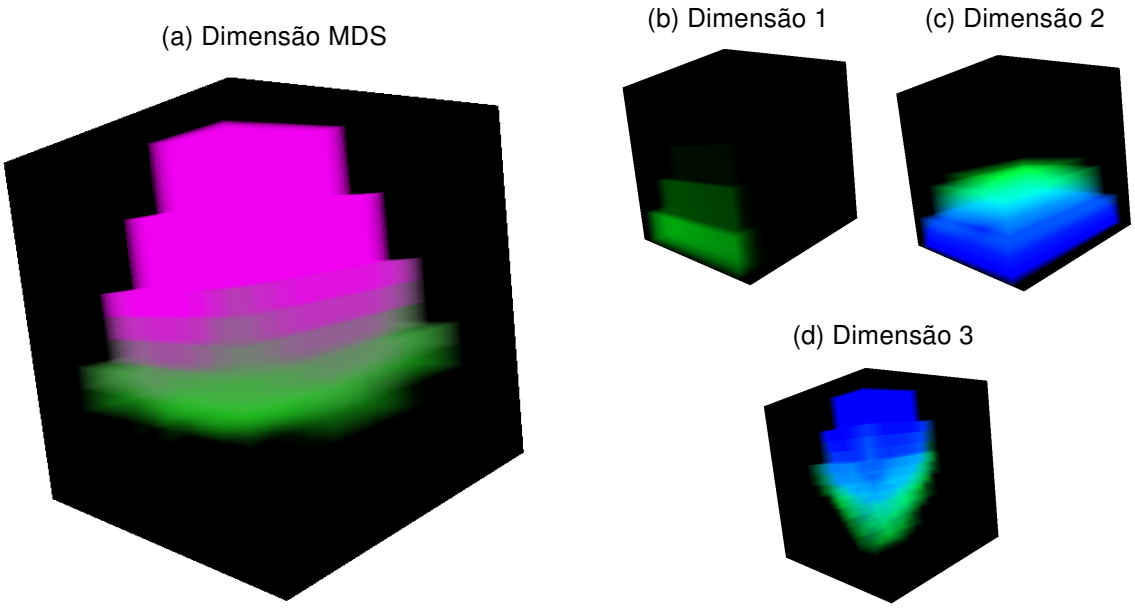
O que corresponderia a uma representação matricial como a da Equação (46) com pelo menos três colunas de matrizes, ou no caso do conteúdo da Figura 22, células com pelo menos 3 valores em cada uma.

O procedimento adotado para projetar um volume em um espaço bidimensional é formado pelos seguintes passos:

1. Na primeira etapa do processo, para otimizar o processamento, considera-se apenas uma instância de cada vetor, ou seja, caso um vetor esteja presente em mais de uma célula da matriz, ele contará com apenas uma instância nas próximas etapas. A posição em que cada um dos vetores ocorre na matriz de entrada é armazenada em uma matriz de mapeamento, e uma nova matriz de dados com elementos singulares é construída;
2. Calcular a matriz de distâncias entre os elementos distintos da matriz de elementos singulares, obtida no passo 1;
3. Realizar a projeção da matriz de elementos singulares para o espaço de destino, com $m = 2$ (plano de cores);
4. Aplicar o algoritmo SMACOF, conforme apresentado na Seção 2.4.2, utilizando a matriz de distâncias do passo 2 e a projeção encontrada no passo 3 como parâmetros de entrada;
5. Transformar os vetores obtidos com a execução do algoritmo SMACOF para o intervalo $[0; 255]$. Esta transformação linear é aplicada para ajustar os resultados aos valores utilizados na escala de cores do processo de visualização;
6. Considerar os resultados do algoritmo SMACOF e utilizar os valores nas posições correspondentes da matriz de mapeamento do passo 1 para construir um novo volume, cujos vetores de dados agora representam as coordenadas em um plano de cores;
7. Utilizar a matriz de dados reduzida, obtida no passo 4 como entrada para o algoritmo de visualização, para a geração da representação do volume de dados com suas dimensões reduzidas.

Para exemplificar o resultado deste algoritmo são apresentados na Figura 25 um volume formado por três dimensões e a sua correspondente projeção em um espaço de cores bidimensional. A Figura 25b corresponde à primeira dimensão da matriz original, a Figura 25c apresenta o conteúdo da segunda dimensão e a Figura 25d apresenta o conteúdo da terceira dimensão. Na Figura 25a tem-se o resultado do algoritmo proposto, apresentando as características combinadas dos dados presentes nas três dimensões do volume original.

FIGURA 25 – Visualização da projeção bidimensional de um volume com três dimensões



FONTE: Autoria própria

Da mesma maneira que a projeção no espaço unidimensional, este procedimento gera um conjunto de valores bidimensionais interpretados como índices para duas escalas de cores que variam no intervalo: $[0; 255]$. As escalas de cores disponíveis também são definidas no espaço de cores RGB, e o processo de geração das imagens utiliza o resultado da redução dimensional como o campo escalar para a resolução da integral de visualização de volumes, definida na Seção 2.1.1, utilizando as características de opacidade escolhidas de forma interativa. Estão disponíveis no protótipo as escalas representadas na Figura 26.

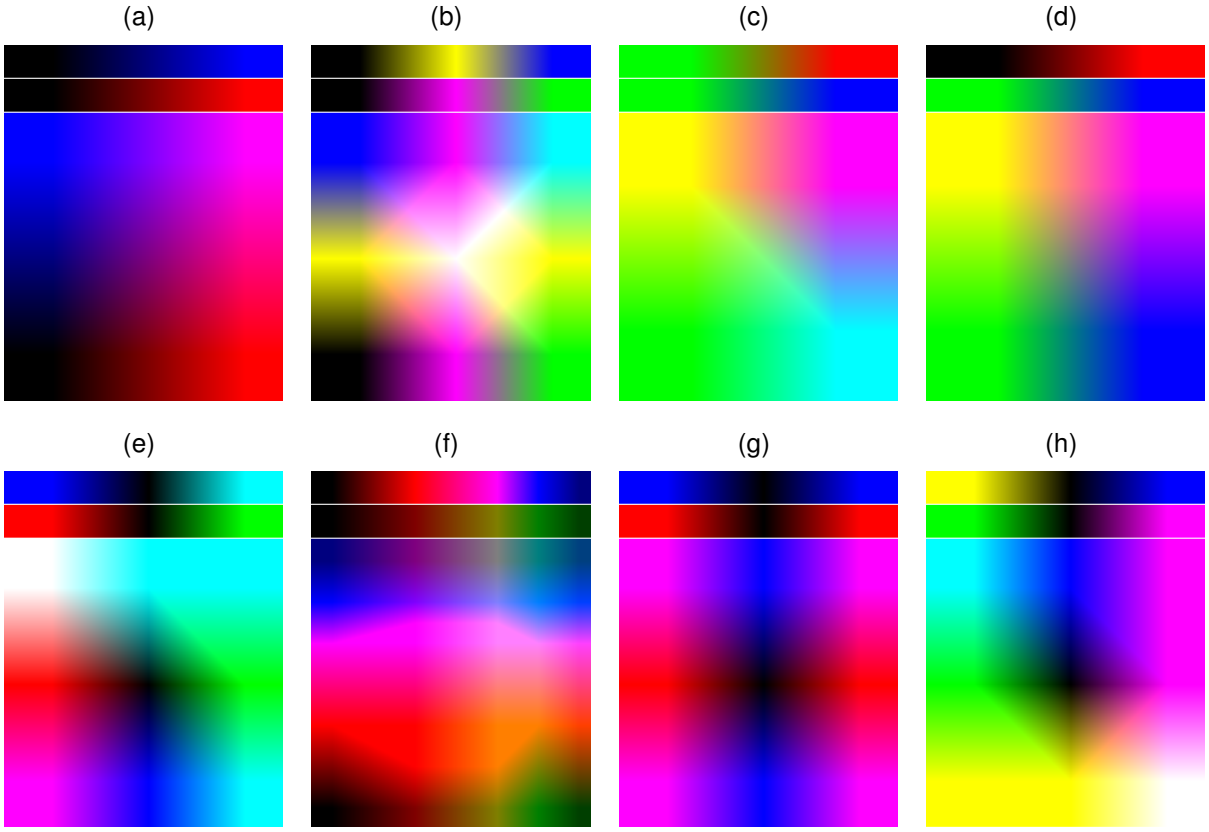
4.3 PROJEÇÕES BIDIMENSIONAIS UTILIZANDO COORDENADAS ESTRELA

No trabalho Matrakas e Scheer (2016b) é definido que o mesmo processo pode ser realizado para a visualização de um conjunto de elementos como o da Figura 25, porém utilizando Coordenadas Estrela para realizar a Redução Dimensional, com a aplicação dos seguintes passos:

1. A primeira etapa do processo consiste em normalizar os valores de cada uma das dimensões para o intervalo $[0; 1]$, de modo que todas as variáveis tenham o mesmo peso no resultado da transformação, conforme apresentado na Seção 2.4.15;

2. Realizar a projeção da matriz de elementos normalizados para o espaço de destino, com $m = 2$ (plano de cores). No caso das Coordenadas Estrela, por definição o espaço de destino é bidimensional, diferente do *Multidimensional Sca-*

FIGURA 26 – Escalas de cores utilizadas nas projeções bidimensionais



FONTE: Autoria própria

ling (MDS) cujo espaço de destino pode ser definido com um número arbitrário de dimensões;

3. Transformar os vetores obtidos com a execução da transformação em Coordenadas Estrela para o intervalo $[0; 255]$. Esta transformação linear é aplicada para ajustar os resultados aos valores utilizados na escala de cores do processo de visualização;

4. Utilizar a matriz de dados normalizada, obtida no passo anterior como entrada para o algoritmo de visualização, para a geração da representação do volume de dados com suas dimensões reduzidas.

Para exemplificar o resultado da projeção em Coordenadas Estrela na Figura 27 está representada a visualização a partir da projeção do mesmo volume de dados da Figura 25.

4.4 PROJEÇÕES BIDIMENSIONAIS UTILIZANDO COORDENADAS ESTRELA E MULTIDIMENSIONAL SCALING

O processo de RD realizado pelo algoritmo SMACOF necessita que uma projeção inicial não nula seja fornecida ao seu processo, de modo que seja realizado o ajuste dos valores desta matriz inicial de forma iterativa, buscando a configuração com o menor valor de Stress, ou seja, minimizando as diferenças entre as distâncias para cada par de pontos no espaço de destino com as distâncias correspondentes ao mesmo par no espaço de origem.

Desta maneira, o processo de redução dimensional utilizando o SMACOF pode ser adaptado para utilizar como projeção inicial a transformação por Coordenadas Estrela, com a aplicação dos seguintes passos:

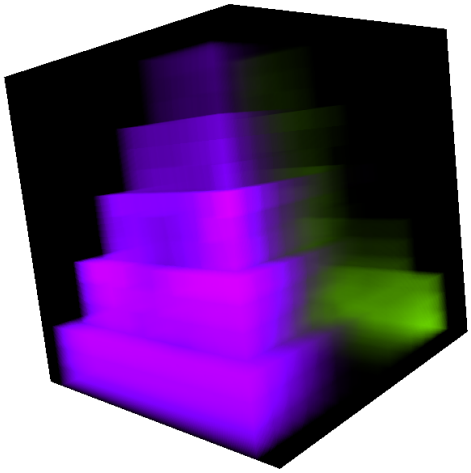
1. A primeira etapa do processo consiste em normalizar os valores de cada uma das dimensões para o intervalo $[0; 1]$, de modo que todas as variáveis tenham o mesmo peso no resultado da transformação;

2. Considera-se apenas uma instância de cada vetor, ou seja, caso um vetor esteja presente em mais de uma célula da matriz, ele contará com apenas uma instância nas próximas etapas. A posição em que cada um dos vetores ocorre na matriz de entrada é armazenada em uma matriz de mapeamento, e uma nova matriz de dados com elementos singulares é construída;

3. Calcular a matriz de distâncias entre os elementos distintos da matriz de elementos singulares, obtida no passo 2;

4. Realizar a projeção da matriz de elementos singulares utilizando Coordenadas Estrela, cujo resultado deve ser considerado como a projeção inicial do

FIGURA 27 – Projeção em Coordenadas Estrela de um volume com três dimensões



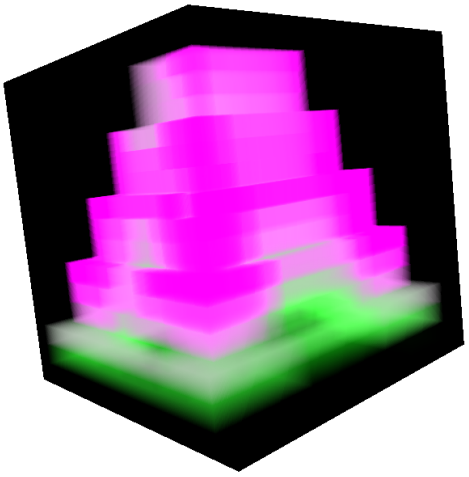
FONTE: Autoria própria

SMACOF;

5. Aplicar o algoritmo SMACOF utilizando a matriz de distâncias do passo 3 e a projeção encontrada no passo 4 como parâmetros de entrada;
6. Transformar os vetores obtidos com a execução do algoritmo SMACOF para o intervalo $[0; 255]$. Esta transformação linear é aplicada para ajustar os resultados aos valores utilizados na escala de cores do processo de visualização;
7. Considerar os resultados do algoritmo SMACOF e utilizar os valores nas posições correspondentes da matriz de mapeamento do passo 2 para construir um novo volume, cujos vetores de dados agora representam as coordenadas em um plano de cores;
8. Utilizar a matriz de dados reduzida, obtida no passo 7 como entrada para o algoritmo de visualização, para a geração da representação do volume de dados com suas dimensões reduzidas.

Para exemplificar o resultado da projeção em Coordenadas Estrela na Figura 28 está representada a visualização a partir da projeção do mesmo volume de dados da Figura 25.

FIGURA 28 – Visualização do volume apresentado na Figura 25 com projeção inicial em Coordenadas Estrela e aplicação do SMACOF



FONTE: Autoria própria

Neste capítulo foram descritas quatro abordagens de utilização de técnicas de RD com o objetivo de criar uma codificação que possa ser utilizada como fonte em um algoritmo de visualização de volumes. As diferenças entre as abordagens dizem respeito ao espaço de destino da RD e ao algoritmo empregado para realizar a RD.

Como parte da solução, foi apresentada uma descrição, juntamente com o seu diagrama correspondente, do fluxo de dados necessário para atingir a imagem representativa dos dados de entrada, bem como algumas considerações a respeito das

características de processamento das técnicas consideradas na implementação do protótipo.

A utilização do protótipo para o processamento das bases de dados descritas no Capítulo 3 produziu um conjunto de resultados que serão apresentados no próximo capítulo.

5 RESULTADOS

Este capítulo apresenta os resultados obtidos durante a execução do trabalho, explicitando as diferentes fases intermediárias da pesquisa e desenvolvimento. São apresentadas as imagens resultantes da execução dos algoritmos descritos no Capítulo 4, quais sejam: redução para espaços unidimensional e bidimensional utilizando o *Scaling by Mojorizing a Complicated Function* (SMACOF), redução para um espaço bidimensional utilizando a transformação para Coordenadas Estrela e também a redução para um espaço bidimensional utilizando a projeção para Coordenadas Estrela como ponto de partida para o algoritmo SMACOF.

Os testes aqui descritos foram realizados utilizando as bases de dados e o ambiente computacional caracterizados no Capítulo 3, estando o texto organizado em seções correspondentes a estes conjuntos de dados.

5.1 PROCESSAMENTO DO BLOCO 22

A matriz correspondente ao Bloco 22 é fornecida aos algoritmos de redução dimensional sem tratamentos adicionais aos seus valores. Estão representadas na Figura 29 as projeções correspondentes aos 982 vetores distintos que formam o bloco artificial nos espaços uni e bidimensional.

As Figuras 29a e 29b mostram as projeções resultantes da execução da Redução Dimensional (RD) utilizando o algoritmo SMACOF, a partir das suas duas configurações consideradas neste trabalho. Na Figura 29c está retratado o resultado obtido a partir da execução da transformação por Coordenadas Estrela. Por último, o conteúdo da Figura 29d reproduz a projeção resultante da aplicação do SMACOF considerando como projeção inicial o resultado obtido com a RD por Coordenadas Estrela.

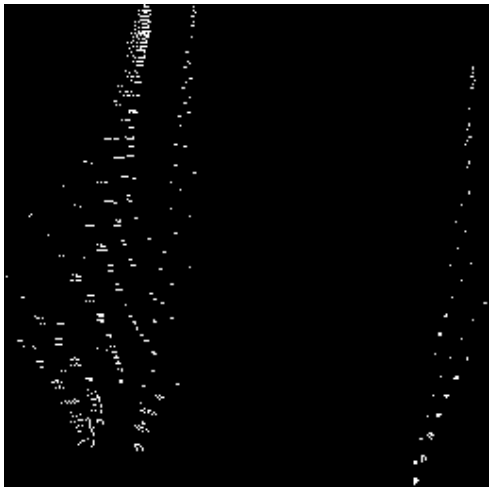
De acordo com a descrição do processo proposto no Capítulo 4, o tratamento realizado nos dados é a normalização aplicada no resultado da RD, na qual os valores são ajustados para o intervalo de números inteiros $[0; 255]$, de maneira a facilitar sua interpretação no espaço de cores utilizado na etapa de visualização do volume. Após a execução da RD, cada vetor do volume de dados original é substituído pelas coordenadas correspondentes à sua projeção no plano de cores, que serão utilizados no processo de geração da imagem correspondente à representação do conteúdo de todas as dimensões dos dados.

O resultado deste processo aplicado ao volume de dados correspondente ao Bloco 22 é apresentado nas Figuras 30, 31, 32 e 33. Para cada um dos algoritmos

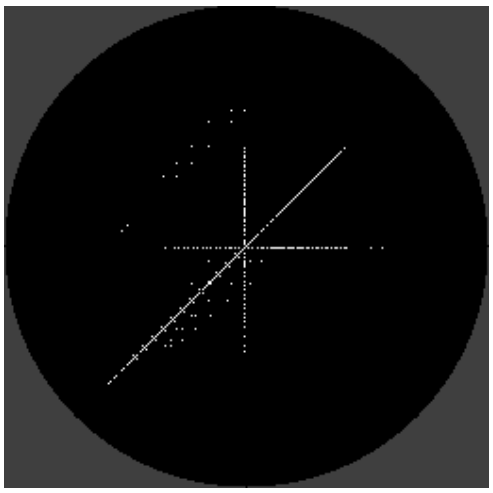
FIGURA 29 – Resultados da redução dimensional utilizando os algoritmos implementados

(a) SMACOF 1D

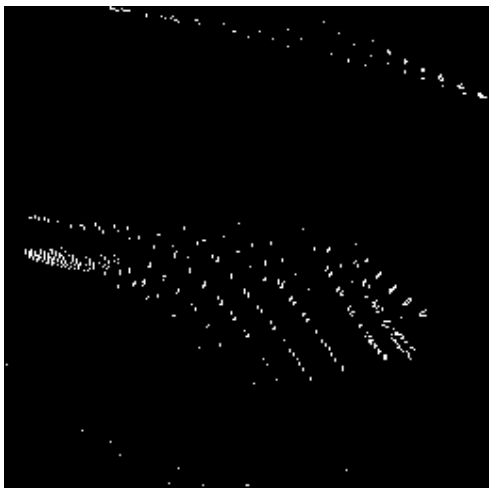
(b) SMACOF 2D



(c) Coordenadas Estrela



(d) SMACOF com Coordenadas Estrela



FONTE: Autoria própria

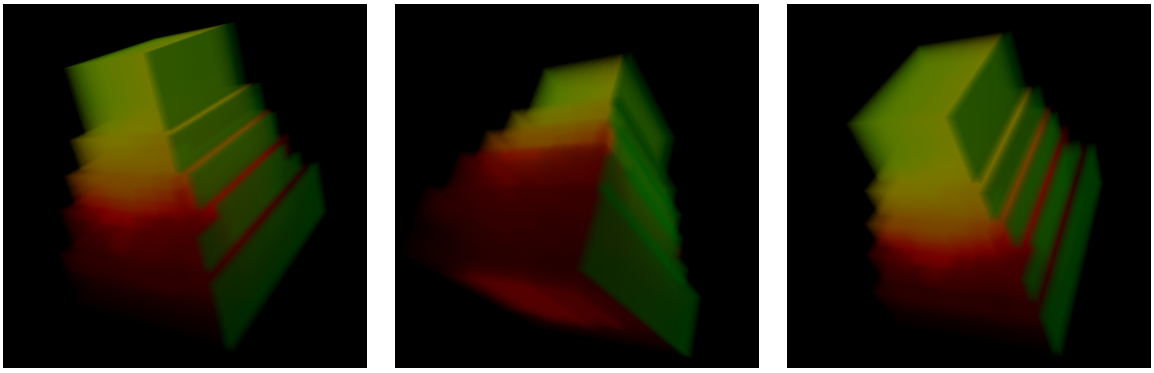
estão representadas imagens com o mesmo ângulo de visualização dos dados apresentados na Figura 13 e também são mostrados mais dois ângulos de maneira que seja possível visualizar a base e o topo do bloco representado no volume de dados.

Como, por definição, a origem do espaço de Coordenadas Estrela ocupa o centro do plano de destino da transformação, a escala de cores a ser utilizada precisa levar esta característica em consideração, com a cor preta ocupando o centro do plano de cores, motivo pelo qual algumas das escalas apresentadas na Figura 26 possuem a cor preta na região central. O mesmo não ocorre com os resultados obtidos com o algoritmo SMACOF, e portanto as escalas que apresentam a cor preta no canto inferior esquerdo são mais adequadas para representar a sua saída.

A cor preta é interpretada pelo algoritmo de visualização como ausência de valores, não contribuindo na imagem gerada para o volume de dados em questão. Isto porque conforme o exposto na Seção 2.2 o preto é representado por componentes de

FIGURA 30 – Renderização da RD com SMACOF 1D para o volume do Bloco 22

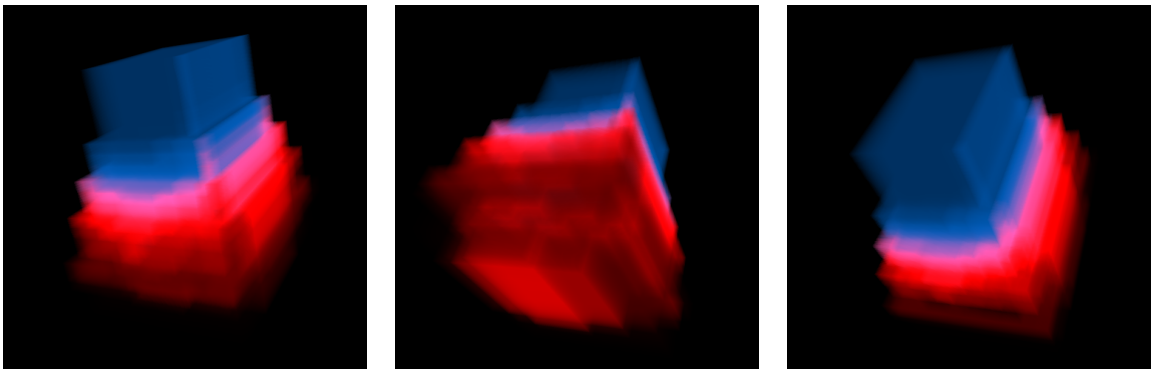
(a) (b) (c)



FONTE: Autoria própria

FIGURA 31 – Renderização da RD com SMACOF 2D para o volume do Bloco 22

(a) (b) (c)



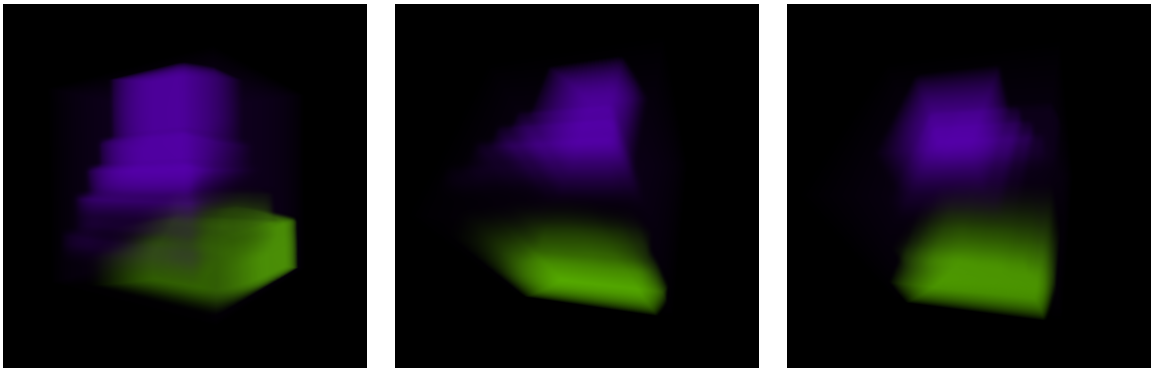
FONTE: Autoria própria

valor 0, que não vão alterar o resultado do cálculo da integral de visualização aplicada ao volume de dados. Esta correspondência entre a cor preta e a ausência de valores na representação é necessária também pelo fato evidenciado na Seção 3.2, que diz que em um volume podem existir regiões nas quais não existem valores a serem representados.

Os resultados visuais obtidos com cada um dos algoritmos apresentam características distintas da matriz de entrada, revelando relacionamentos entre os vetores de acordo com as especificidades de cada um. As imagens produzidas a partir das projeções calculadas com o algoritmo SMACOF, mostradas nas Figuras 30 e 31, tanto uni quanto bidimensional, evidenciam de forma mais expressiva o conteúdo da quarta dimensão do conjunto original de dados, representada na Figura 13d. O mesmo não ocorre nas representações produzidas a partir dos resultados obtidos com a utilização da RD por Coordenadas Estrela, conteúdo das Figuras 32 e 33.

FIGURA 32 – Visualização do resultado da RD com Coordenadas Estrela para o volume do Bloco 22

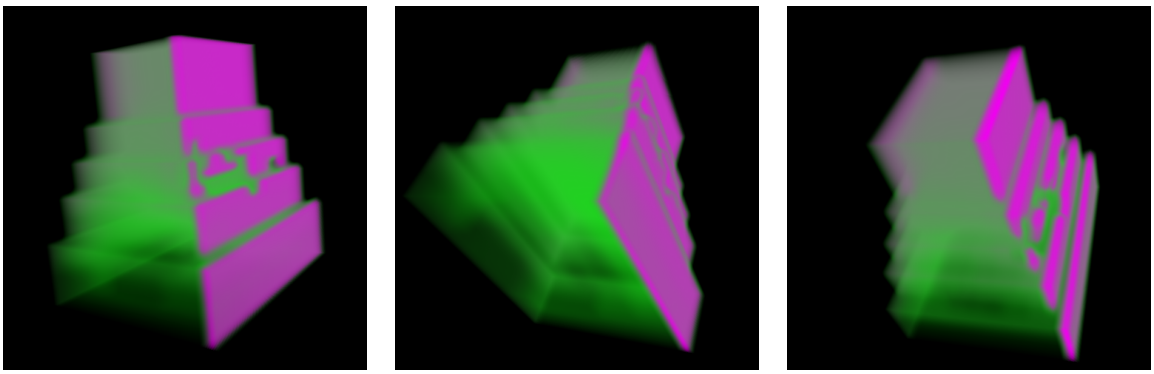
(a) (b) (c)



FONTE: Autoria própria

FIGURA 33 – Visualização do resultado da RD com Coordenadas Estrela e SMACOF 2D para o volume do Bloco 22

(a) (b) (c)



FONTE: Autoria própria

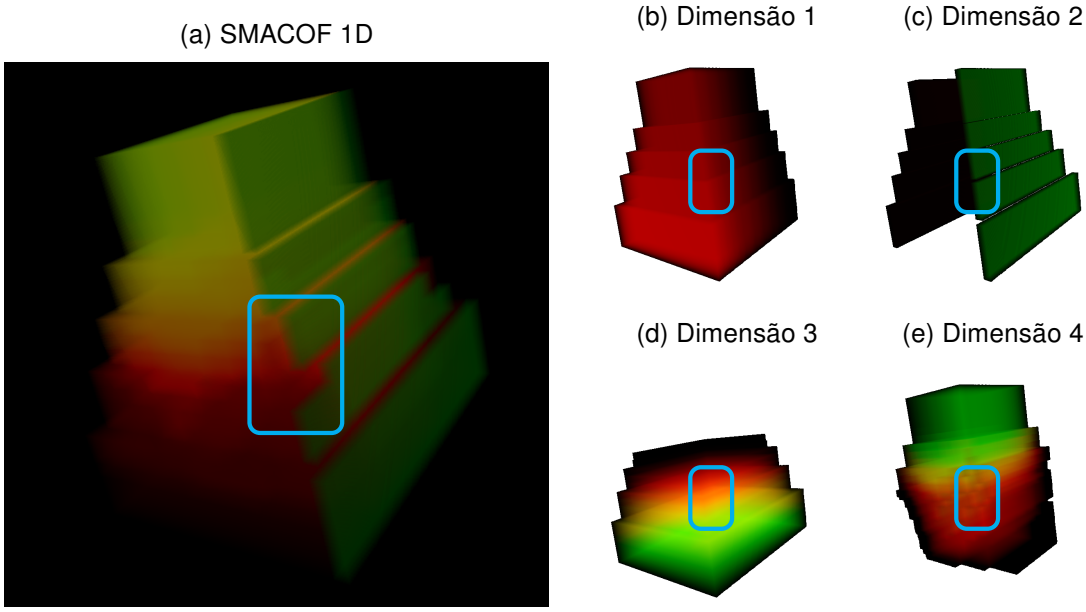
5.1.1 Resultados do processamento do SMACOF unidimensional

De modo a caracterizar melhor as diferenças visuais nos resultados obtidos, na Figura 34 estão destacadas as regiões nas quais estão presentes relações entre os dados que não ficam evidentes no conjunto de dados originais, reproduzidos com a mesma região destacada nas Figuras 34b, 34c, 34d e 34e, sendo que todas as variáveis representadas utilizam o mapa de cores apresentado na Figura 24c. Na área marcada na Figura 34a também é possível observar a influencia da dimensão 4, Figura 34e, na resposta.

5.1.2 Resultados do processamento do SMACOF bidimensional

Nos resultados obtidos com a projeção bidimensional ao executar o algoritmo SMACOF, de forma semelhante ao ocorrido no caso unidimensional, a dimensão 4 tem uma influência dominante na representação visual resultante, destacada na Figura 35, com a mesma área evidenciada nas imagens correspondentes às dimensões

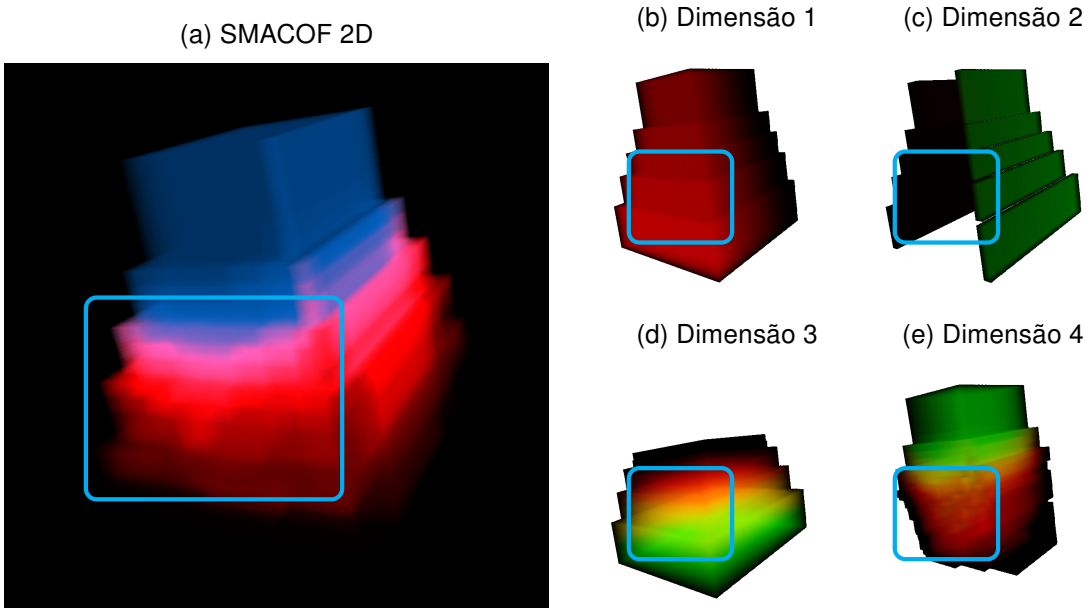
FIGURA 34 – Detalhe da RD com SMACOF 1D para o volume do Bloco 22



FONTE: Autoria própria

presentes no conjunto de dados original, reproduzidas nas Figuras 35b, 35c, 35d e 35e, e de maneira análoga à discussão do resultado para SMACOF unidimensional, todas as variáveis representadas utilizam o mapa de cores apresentado na Figura 24c. Neste caso, não existe uma região que represente uma alteração na continuidade dos conteúdos observáveis dos dados originais, como o destacado na Figura 34a, mas sim a ausência de referência visual para a dimensão 2 correspondente ao conteúdo da Figura 34c.

FIGURA 35 – Detalhe da RD com SMACOF 2D para o volume do Bloco 22



FONTE: Autoria própria

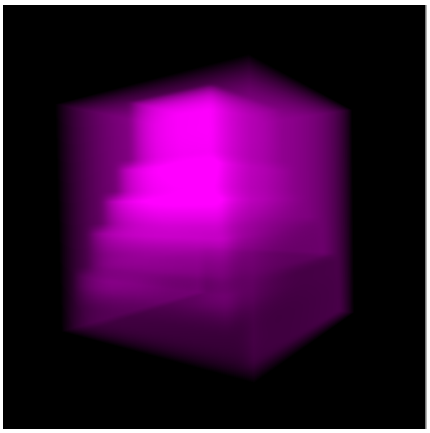
5.1.3 Resultados do processamento da transformação por Coordenadas Estrela

Diferentemente dos resultados apresentados pelas duas configurações do algoritmo SMACOF, a RD por Coordenadas Estrela não tem uma correspondência direta com as representações dos campos escalares que compõem o conjunto de dados original. Esta é uma característica resultante do tipo de transformação realizada nos vetores de dados.

No caso do *Multidimensional Scaling* (MDS), de acordo com Borg e Groenen (2005), a transformação resulta em um espaço sobre o qual não se tem alguma informação a priori, e cujos vetores são normalizados conforme os algoritmos apresentado nas Seções 4.1 e 4.2, resultando em um conjunto resposta pertencente ao primeiro quadrante do plano cartesiano, no qual é realizado o mapeamento com o plano de cores utilizado para criar a representação visual dos dados originais. Já no caso das Coordenadas Estrela, os eixos do espaço original são dispostos de forma circular no plano cartesiano, levando a uma configuração na qual a origem do espaço *n*-dimensional vai ocupar sempre o centro do conjunto resultante de pontos projetados no plano de cores, razão pela qual o vetor nulo, representado pela cor preta, deve estar no centro do plano de cores neste caso.

A diferença entre estas representações está evidenciada na Figura 36, que mostra o resultado do processamento do Bloco 22 por Coordenadas Estrela e utilizando o mesmo plano de cores empregado ao gerar as imagens de resposta para os resultados obtidos com o SMACOF bidimensional. O conteúdo da Figura 36 é o mesmo daquele apresentado na Figura 32, porém é possível ver a representação de valores em regiões nas quais não existem dados, ou seja, onde deveria ser representada uma região vazia do volume de dados.

FIGURA 36 – Resultado da RD por Coordenadas Estrela utilizando plano de cores do MDS

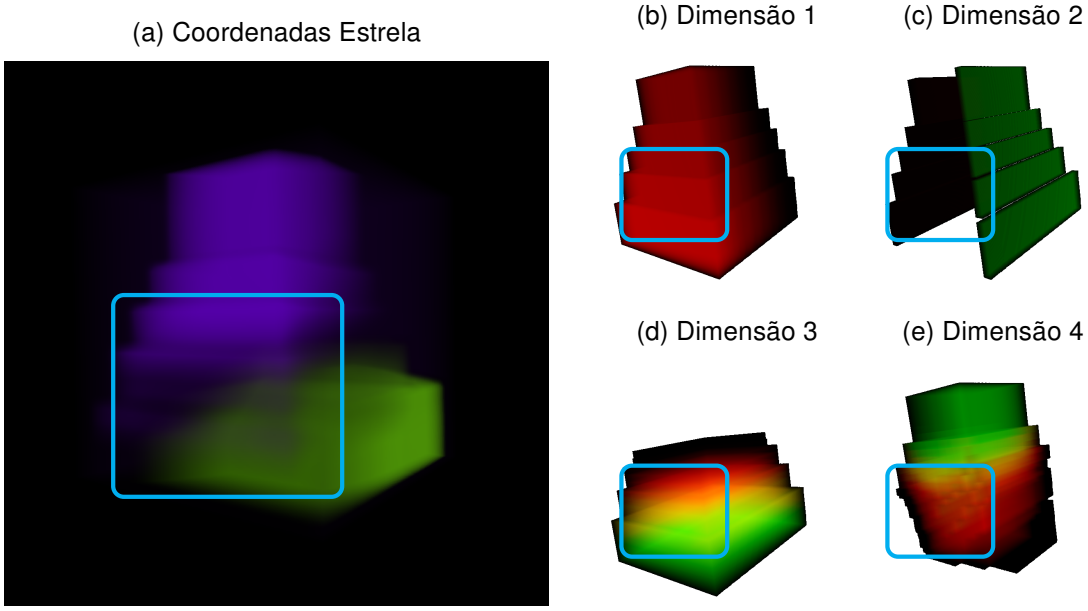


FONTE: Autoria própria

As imagens resultantes da RD, ao executar a transformação por Coordenadas

Estrelas, não têm uma correspondência direta com as representações das variáveis originais do conjunto de dados, como evidenciado pela área delimitada na Figura 37a que contém uma região semitransparente. Nas Figuras 37b, 37c, 37d e 37e não existem evidências visuais que equivalem ao conteúdo denotado na Figura 37a, e novamente todas as variáveis representadas utilizam a mesma escala de cores, que está representada na Figura 38.

FIGURA 37 – Detalhe da RD por Coordenadas Estrela para o volume da Figura 13



FONTE: Autoria própria

FIGURA 38 – (a) Escala de cores utilizada na projeção da Figura 37a (b) Escala de cores utilizada nas projeções das Figuras 37b, 37c, 37d e 37e

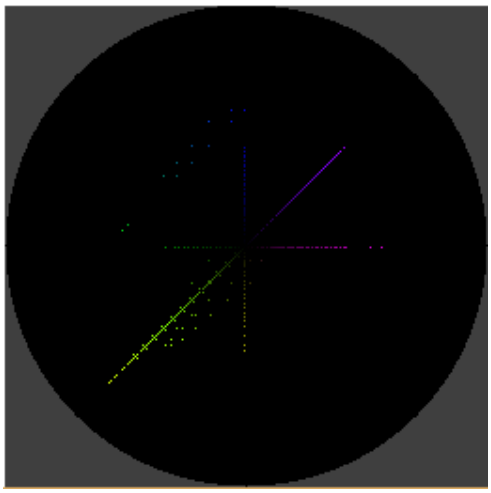
FONTE: Autoria própria

A região semitransparente do volume obtido como resposta é composta por valores próximos de zero, o que leva à conclusão que os vetores que a compõem são mapeados próximos da origem do sistema de Coordenadas Estrela. Esta correspondência entre valores projetados e a sua representação visual está demostrada na Figura 39, na qual os vetores da matriz de entrada estão representados no plano de Coordenas Estrela utilizando a codificação de cores utilizado pelo algoritmo de visualização ao criar a representação do volume mostrado na Figura 37a.

5.1.4 Resultados do processamento do SMACOF com Coordenadas Estrela

Conforme apresentado na Seção 4.4, o algoritmo SMACOF inicia seu processo iterativo a partir do resultado da transformação por Coordenadas Estrela do volume

FIGURA 39 – Pontos projetados com Coordenadas Estrela no plano de cores

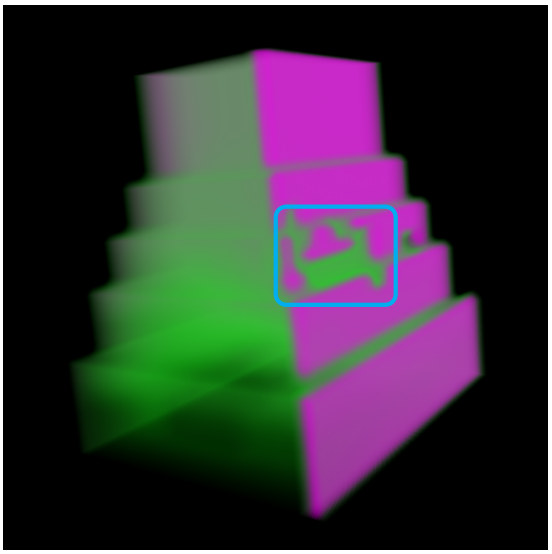


FONTE: Autoria própria

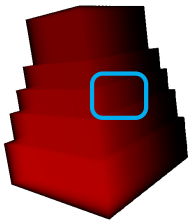
original, resultando na imagem retratada na Figura 40. Semelhantemente ao resultado obtido com projeção unidimensional utilizando o SMACOF e também com o resultado alcançado com a transformação por Coordenadas Estrela, as imagens obtidas com este algoritmo apresentam características distintas da matriz de entrada, revelando relacionamentos que não aparecem ao se analisar visualmente as variáveis originais. Estas características estão destacadas na Figura 40a, com a mesma área evidenciada nas imagens correspondentes dimensões presentes no conjunto de dados original, reproduzidas nas Figuras 40b, 40c, 40d e 40e, e mais uma vez, todas as variáveis representadas utilizam o mapa de cores apresentado na Figura 24c.

FIGURA 40 – Detalhe da RD por SMACOF com projeção inicial a partir de Coordenadas Estrela para o volume da Figura 13

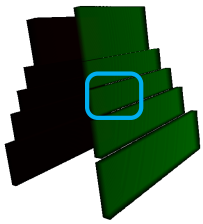
(a) SMACOF com Coordenadas Estrela



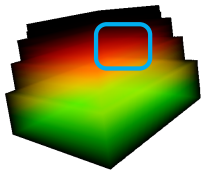
(b) Dimensão 1



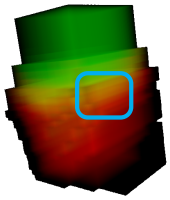
(c) Dimensão 2



(d) Dimensão 3



(e) Dimensão 4



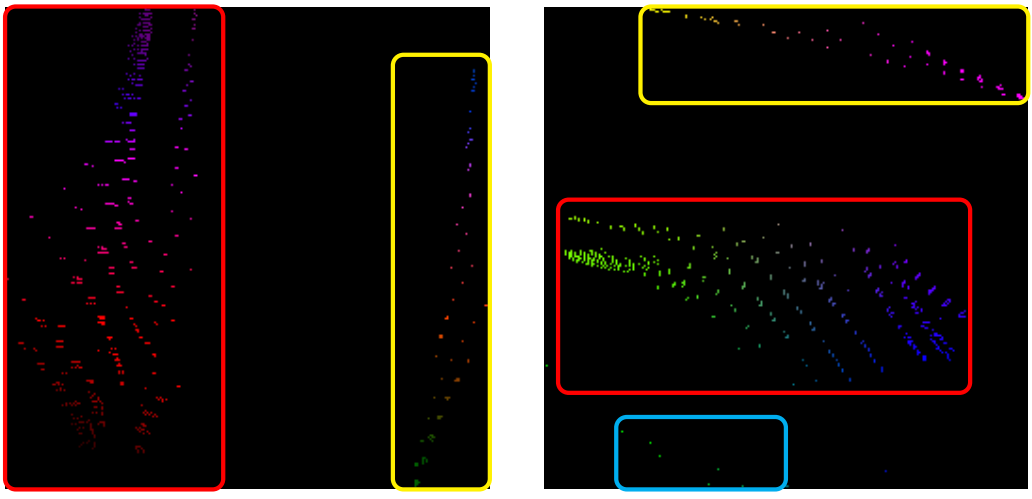
FONTE: Autoria própria

Os resultados obtidos a partir da execução do SMACOF, utilizando a projeção inicial produzida pela transformação por Coordenadas Estrela, não evidenciam no resultado o conteúdo da dimensão 4, representado na Figura 40e, como no caso do SMACOF bidimensional, apresentado na Seção 5.1.2. Em contrapartida, o conteúdo da dimensão 2, representado na Figura 40c, está em evidência na imagem resultante, o que leva à conclusão que a projeção inicial tem grande influência na configuração final da resposta produzida no processo iterativo.

Estas diferenças de representação também aparecem na disposição das projeções dos vetores no plano de cores, representadas nas Figuras 29b e 29d. Na Figura 41 são reproduzidas as imagens que correspondem e estas projeções nos planos de cores utilizados para a produção das Figuras 31 e 33. Estão destacados nas Figuras 41a e 41b os grupos de vetores formados, considerando que ao utilizar a transformação por Coordenadas Estrela, o espaço resultante apresenta uma rotação em relação à orientação apresentada no resultado do SMACOF. Além disso, existe também a formação de um novo grupo, que está destacado na cor ciano. Com relação aos grupos marcados em vermelho nas duas projeções, existe uma reconfiguração da distribuição dos vetores, não apenas a sua rotação, alterando os relacionamentos existentes entre os seus elementos.

FIGURA 41 – Comparação entre resultados da RD utilizando: (a) SMACOF (b) SMACOF com Coordenadas Estrela

(a) SMACOF 2D (b) SMACOF com Coordenadas Estrela



FONTE: Autoria própria

É conveniente destacar que os grupos marcados nas projeções da Figura 41 são o resultado da sua inspeção visual, com o objetivo de destacar as diferenças nos resultados obtidos com a execução dos dois algoritmos.

5.1.5 Características da execução dos algoritmos para o Bloco 22

O algoritmo SMACOF, por ser um processo numérico iterativo, necessita verificar as condições correspondentes ao valor da função de STRESS e, caso um determinado limite de precisão seja atingido, uma aproximação aceitável para a configuração de distribuição das distâncias entre os vetores projetados foi atingida. A precisão adotada na implementação do SMACOF neste trabalho foi de 1, determinado após a realização do processamento dos conjuntos de teste artificialmente elaborados, como o Bloco 22 descrito na Seção 3.2.

Uma segunda condição utilizada corresponde ao número de iterações realizadas pelo SMACOF ao processar um conjunto de dados, sendo adotado o limite de 500 iterações. Dentre todos os conjuntos de dados testados, em nenhum caso o limite de iterações foi atingido, obtendo-se uma aproximação adequada com relação à precisão no cálculo da função de STRESS.

Na Tabela 2 estão listados: o número de passos calculados ao processar a matriz correspondente ao Bloco 22 por cada um dos algoritmos. A transformação por Coordenadas Estrela não é apresentada nesta tabela devido ao fato de que os cálculos realizados não são iterativos, mas sim apenas a aplicação de uma única transformação nos vetores de entrada, conforme o disposto na Seção 2.4.15. Além das iterações, são apresentados também, na mesma tabela, os valores iniciais e finais da função STRESS, para cada uma implementações.

TABELA 2 – Número de iterações na execução do SMACOF considerando o volume de dados do Bloco 22

Algoritmo	Passos	Stess inicial	Stress final
SMACOF 1D	17	12.580.603.904	2.029.975.808
SMACOF 2D	69	8.188.079.104	106.124.824
SMACOF com Coordenadas Estrela	293	16.211.548.160	254.745.216

O processo de RD a partir do algoritmo SMACOF consome uma grande quantidade de tempo, tanto no uso da *Central Processing Unit* (CPU) quanto da *Graphical Processing Unit* (GPU), ao serem produzidas as imagens apresentadas como resultado do processamento do Bloco 22. O mesmo não pode ser afirmado para o processamento da transformação por Coordenadas Estrela, que é executada em aproximadamente 5ms, considerando o mesmo conjunto de dados como entrada. Os tempos de execução para cada um dos algoritmos implementados está explicitado na Tabela 3.

Terminada a etapa de RD é iniciado o processo interativo do protótipo, no qual a representação do volume pode ser manipulada pelo usuário, sendo possível alterar os parâmetros de brilho, escala e deslocamento da função de transferência, o plano de cores, empregado para mapear os valores existentes na matriz de dados nas cores da representação visual desta matriz. É possível também alterar a densidade aplicada

TABELA 3 – Tempos aproximados de execução dos algoritmos sobre o volume de dados do Bloco 22

Algoritmo	**Tempo Total (ms)**	**Tempo iteração (ms)**
SMACOF 1D	313	18
SMACOF 2D	899	13
SMACOF com Coordenadas Estrela	6094	20
Coordenadas Estrela	5	*não se aplica*

aos valores da matriz de entrada.

No ambiente computacional descrito na Seção 3.1, a taxa de atualização do resultado visual é de aproximadamente 65 frames por segundo, ou seja, a matriz resultante da RD é transformada um uma imagem a cada 15ms aproximadamente. O tempo de processamento para gerar a representação visual da matriz e o tempo para calcular um passo do algoritmo SMACOF são aproximadamente os mesmos. Tanto o tempo de processamento dos passos do SMACOF quanto o tempo necessário para gerar uma imagem a partir da matriz de dados dizem respeito ao tempo efetivo de processamento na GPU, ou na CPU para o caso da transformação por Coordenadas Estrela, não sendo contabilizadas outras atividades, como a transferência de dados entre a memória de vídeo e a memória principal do computador.

5.2 TESTES COM OS DADOS DE SIMULAÇÃO DO FURACÃO ISABEL

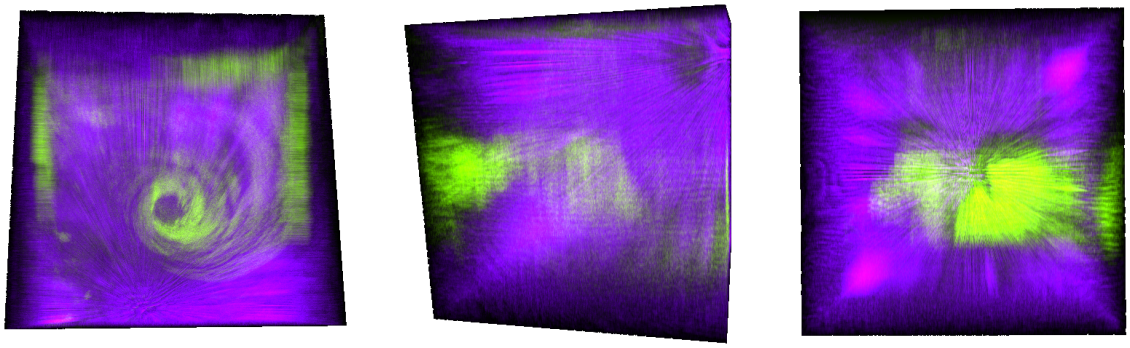
De acordo ao exposto sobre o algoritmo SMACOF na Seção 2.4.2, cujas características incluem restrições de memória, a redução dimensional com este algoritmo não foi possível de ser aplicada no conjunto completo de dados correspondente à simulação do Furacão Isabel, devido ao seu tamanho. Considerando a matriz de dados de entrada, com $[500 \times 500 \times 100]$ elementos, e visto que para cada elemento foram incluídas 8 das 13 dimensões disponíveis no conjunto original de dados, são necessários aproximadamente 763 MB de memória para armazenar a matriz correspondente a um intervalo de tempo. Esta matriz possui um total de 25 milhões de elementos, o que excede em muito o limite de alocação da plataforma computacional sendo utilizada para a execução do algoritmo SMACOF, que suporta o processamento de aproximadamente 13.000 elementos, sendo este o motivo pelo qual são apresentados resultados utilizando a matriz completa para um instante de tempo da simulação do Furacão Isabel, apenas com o algoritmo de RD por Coordenadas Estrela.

Ao aplicar o processo de visualização a partir da RD obtida com a transformação por Coordenadas Estrela, são obtidas imagens como as apresentadas na Figura 42. É possível observar nas imagens algumas das características do furacão, como seu vórtice, formações de nuvens, além da influência do perfil de pressão e distribuição das demais variáveis. O conteúdo da Figura 42 é o resultado da RD a partir das seguintes variáveis da base de dados do Furacão Isabel: Pressão (P), Água

em nuvem (QCLOUD), Graupel (QGRAUP), Gelo em nuvem (QICE), Chuva (QRAIN), Neve (QSNOW), Vapor de água (QVAPOR) e Temperatura (T). Para o resultado em questão, foram considerados os valores para as variáveis indicadas no intervalo de tempo de número trinta.

FIGURA 42 – Projeção Simultânea de 8 variáveis, ou dimensões, do conjunto de dados do furacão isabel, tempo 30.

(a) (b) (c)



FONTE: Autoria própria

A Figura 42a representa uma visão superior, na qual é possível observar o vórtice do furacão, já nas Figuras 42b e 42c, é possível observar a influência do perfil de pressão e das nuvens que compõem o fenômeno.

Além da mudança do ângulo de visão, o controle do brilho e da densidade dos dados no algoritmo de geração das imagens possibilita a busca por características de interação entre as variáveis, que são de difícil visualização quando são projetadas individualmente.

Uma outra ferramenta muito útil na exploração do volume de dados é a possibilidade de alterar as cores utilizadas no plano de destino da RD, conforme evidenciado nos conteúdos da Figura 26. As imagens apresentadas nas Figuras 16, 17 e 42 utilizam ainda outras 3 escalas de cores, mais adequadas ao conteúdo de cada uma das situações.

O tempo de processamento do algoritmo de transformação por Coordenadas Estrela na matriz de dados do intervalo número trinta da simulação do Furacão Isabel é de 4.102ms. Como nenhum dos demais algoritmos implementados pôde ser utilizado com este conjunto de dados, não há dados comparativos.

5.2.1 Processamento de uma região da matriz de dados do Furacão Isabel

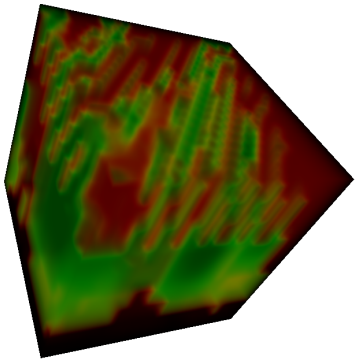
Apesar de não ser possível processar a matriz completa de um intervalo de tempo desta simulação, uma fração do seu volume pode ser submetido às diferentes implementações do algoritmo SMACOF disponíveis. A matriz de ordem $[27 \times 20 \times 35]$,

correspondente ao recorte descrito na Seção 3.3.1, é fornecida aos algoritmos de RD sem tratamentos adicionais aos valores dos seus 18.900 vetores, dos quais 13.198 são distintos.

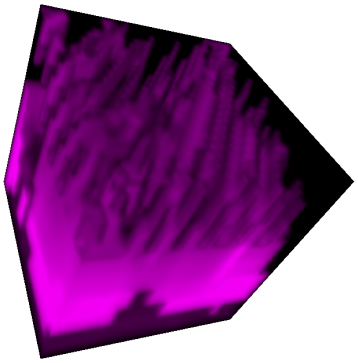
Após a normalização do resultado dos algoritmos de RD aplicados no volume resultante do recorte, as respostas visuais para cada uma das técnicas implementadas são apresentadas na Figura 43. Para cada um dos algoritmos está representada uma imagem com o mesmo ângulo de visualização das variáveis apresentadas na Figura 19.

FIGURA 43 – Imagens produzidas a partir do volume recortado da simulação do Furacão Isabel

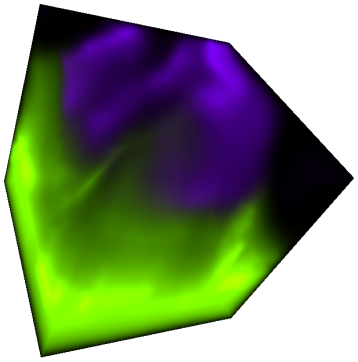
(a) SMACOF 1D



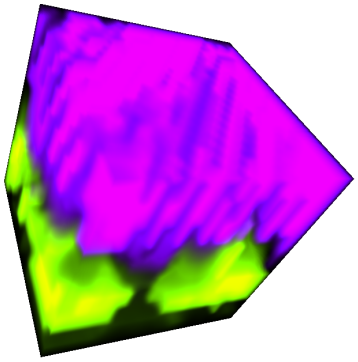
(b) SMACOF 2D



(c) Coordenadas Estrela



(d) SMACOF com Coordenadas Estrela



FONTE: Autoria própria

É possível observar nas imagens características como linhas que correspondem a formações de nuvens, perfis de distribuição vertical dos elementos que compõem o sistema atmosférico em questão. O conteúdo da Figura 19 é o resultado da RD a partir das seguintes variáveis da base de dados do Furacão Isabel: Graupel (QGRAUP), Gelo em nuvem (QICE), Chuva (QRAIN), Neve (QSNOW).

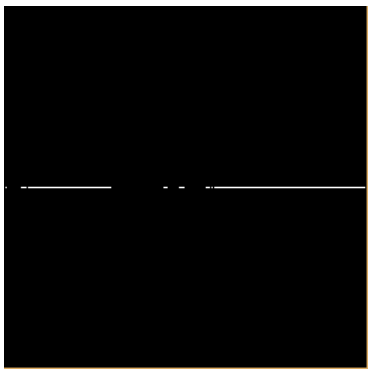
As projeções dos vetores que compõem a matriz com o recorte do conjunto de dados do Furacão Isabel, obtidas a partir da execução de cada um dos algoritmos, estão identificadas na Figura 44.

O conteúdo das Figuras 44a e 44b correspondem aos resultados das duas

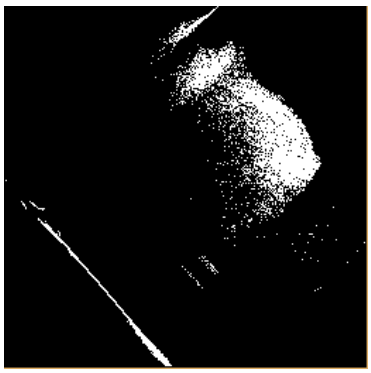
configurações do algoritmo SMACOF, na Figura 44c está retratada a transformação por Coordenadas Estrela, e por último, a Figura 44d mostra a projeção resultante da aplicação do SMACOF com a projeção inicial a partir da transformação por Coordenadas Estrela.

FIGURA 44 – Projeções do volume de recorte da simulação do Furacão Isabel

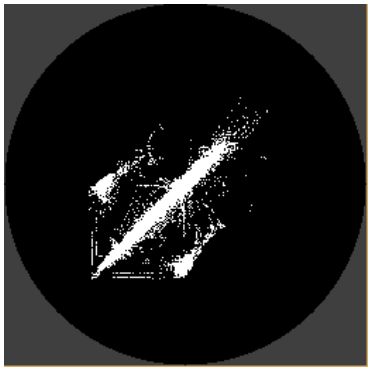
(a) SMACOF 1D



(b) SMACOF 2D



(c) Coordenadas Estrela



(d) SMACOF com Coordenadas Estrela

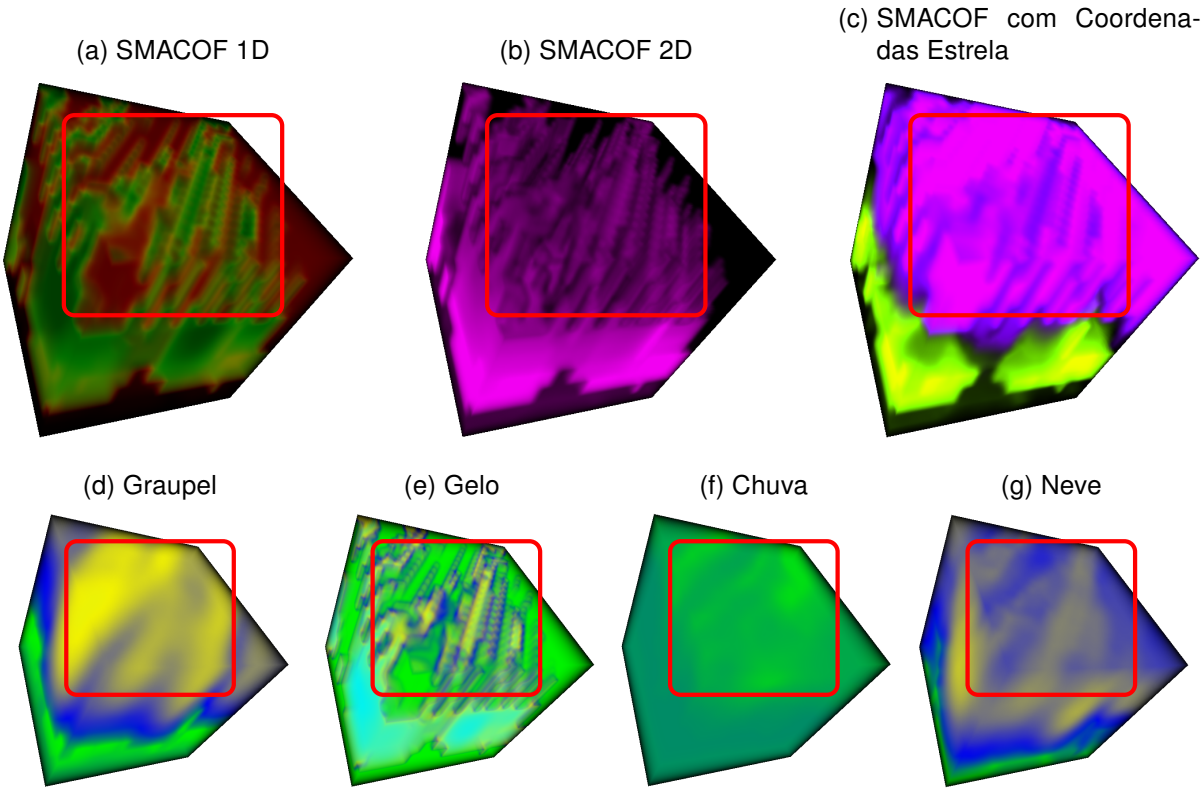
FONTE: Autoria própria

Para caracterizar as diferenças visuais nos resultados obtidos, nas Figuras 45a, 45b e 45c estão destacadas as regiões nas quais estão presentes relações entre os dados que, diferentemente do caso apresentado para o Bloco 22, estão evidentes no conjunto de dados originais, reproduzidos com a mesma região destacada nas Figuras 45d, 45e, 45f e 45g, sendo que todas as variáveis representadas utilizam o mapa de cores apresentado na Figura 24c.

Para o conjunto de teste em questão, as respostas obtidas com todas as diferentes implementações do algoritmo SMACOF destacam as mesmas formações, com variações nos valores escalares resultantes, denotadas pelas configurações nos padrões de cores, porém, sem uma grande diferença na disposição geométrica dos valores.

Diferentemente dos resultados apresentados pelas duas configurações do algoritmo SMACOF, a RD por Coordenadas Estrela novamente não apresenta uma cor-

FIGURA 45 – Detalhe da RD pelas variantes do SMACOF no recorte dos dados do Furacão Isabel



FONTE: Autoria própria

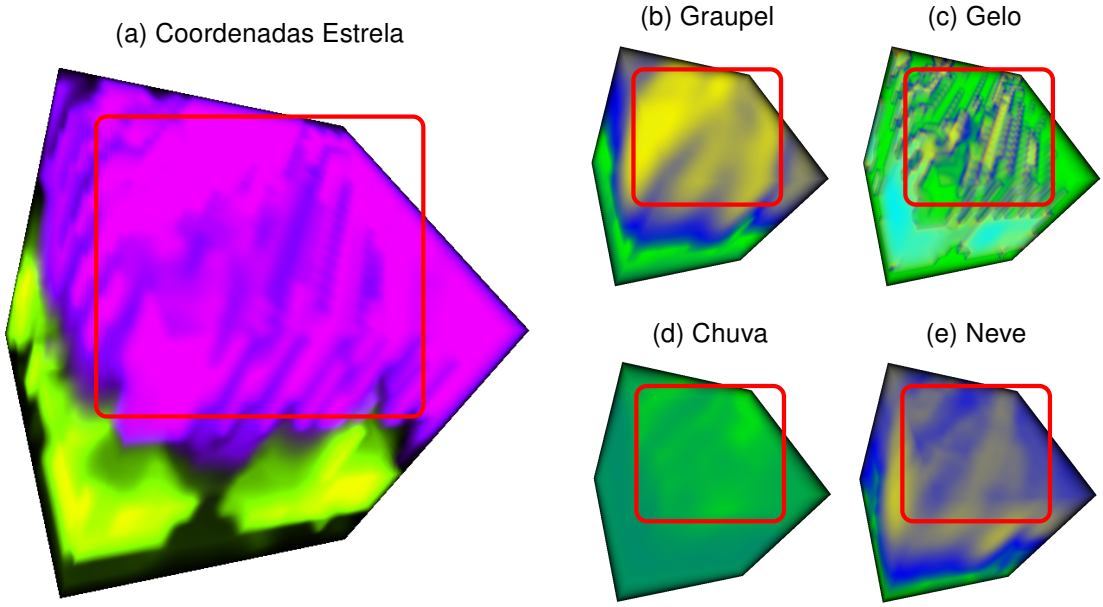
respondência direta com as representações dos campos escalares que compõem o conjunto de dados original.

A diferença entre estas representações pode ser averiguada também na Figura 44, que apresenta as projeções resultantes do processamento para o recorte do conjunto de dados do Furacão Isabel. O grupo de elementos cujas projeções ocupam o centro do plano de cores tem parte de seus componentes não representados na visualização do volume, devido às características do processo e da interpretação dos resultados desta transformação. Em decorrência destes fatores, o resultado obtido apresenta uma região semitransparente marcada na Figura 46a. Nas Figuras 46b, 46c, 46d e 46e não existem evidências visuais que equivalem a este conteúdo, estando a mesma região evidenciada para comparação.

Similarmente ao exposto para o processamento do Bloco 22, na Tabela 4 estão listados para cada dos algoritmos baseados no SMACOF: o número de passos calculados ao processar a matriz correspondente ao recorte do conjunto de dados para o tempo 30 da simulação do Furacão Isabel. Além das iterações, são apresentados também, na mesma tabela, os valores iniciais e finais da função STRESS, para cada uma implementações.

A Tabela 5 apresenta os tempos de processamento que cada uma das técnicas

FIGURA 46 – Detalhe da RD pela transformação por Coordenadas Estrela no recorte dos dados do Furacão Isabel



FONTE: Autoria própria

TABELA 4 – Resultados da execução dos algoritmos sobre o recorte dos dados da simulação do Furacão Isabel

Algoritmo	Passos	Stess inicial	Stress final
SMACOF 1D	39	551.737.950.208	49.017.643.008
SMACOF 2D	58	258.132.590.592	1.392.438.144
SMACOF com Coordenadas Estrela	122	812.925.911.040	704.304.704

disponíveis levou para processar o recorte dos dados proveniente da simulação do Furacão Isabel.

TABELA 5 – Tempos aproximados de execução dos algoritmos sobre o recorte dos dados da simulação do Furacão Isabel

Algoritmo	Tempo Total (ms)	Tempo por iteração (ms)
SMACOF 1D	48.267	1.237
SMACOF 2D	73.636	1.248
SMACOF com Coordenadas Estrela	148.856	1.220
Coordenadas Estrela	16	*não se aplica*

Este capítulo apresentou os resultados obtidos a partir dos testes realizados nos conjuntos de dados, tanto artificialmente gerados para este trabalho, quanto com o resultado da simulação do Furacão Isabel, descritos no Capítulo 3, evidenciando as diferenças das projeções no plano de cores, para cada um dos algoritmos de RD implementados, e as suas consequentes diferenças na representação dos volumes de dados. Destacando também os dados a cerca do desempenho, tanto com relação ao tempo de processamento das bases de dados, como da otimização da função de erro obtida em cada caso.

Estas informações servem de base para as considerações finais a respeito da pesquisa, apresentadas no próximo capítulo.

6 CONSIDERAÇÕES FINAIS

Neste capítulo são apresentados comentários a respeito dos resultados obtidos durante a execução do trabalho, considerando as metas e objetivos estabelecidos na Seção 1.2. Ao final, em uma seção própria, são listados os itens e componentes do projeto que necessitam melhorias, devendo ser abordados na continuação deste, ou de outros trabalhos, que venham a aprimorar as soluções aqui apresentadas.

A proposta deste trabalho foi de apresentar uma abordagem que permita a redução de dimensões e preparação de dados multidimensionais para a sua visualização e análise interativas, utilizando processamento paralelo em um dispositivo computacional de baixo custo. De acordo com o resultado do levantamento bibliográfico apresentado no Capítulo 2, não foram encontrados métodos ou sistemas para realizar a visualização de volumes multidimensionais. As técnicas estudadas, em sua maioria, criam uma representação que pode ser uni, bi ou tridimensional a partir dos dados de entrada, mas sem manter uma relação com a sua disposição espacial, característica que está presente no processo descrito no Capítulo 4.

No Capítulo 4 apresentou-se um diagrama na Figura 21 com as etapas de processamento, de modo que a matriz original de dados seja transformada em uma representação visual do volume de dados que representa. Também foram descritas quatro abordagens para realizar a etapa de Redução Dimensional (RD), necessárias para preparar as matrizes de dados, de modo que o algoritmo de visualização possa gerar as imagens representando o volume de entrada.

No Capítulo 5 são descritas das imagens obtidas a partir destes algoritmos, bem como a apresentadas as características mais relevantes com relação ao número de iterações e tempos de processamento destes algoritmos. A partir dos dados e imagens apresentados, é possível afirmar que o objetivo principal do trabalho foi cumprido, com a definição de procedimentos para preparar e processar volumes de dados, bem como o levantamento de características e restrições de uso destes procedimentos, em um equipamento computacional de baixo custo.

Conforme descrito no Capítulo 4, a interpretação do resultado dos algoritmos de RD como um plano de cores, e no caso unidimensional uma escala de cores, permite a visualização do volume de dados contendo a influência de todas as suas dimensões, ou variáveis. As imagens resultantes deste processo caracterizam uma nova ferramenta para a análise interativa de dados, possibilitando assim a investigação de novas características em dados volumétricos de forma visual.

Apesar de a interpretação das representações obtidas não estar no escopo do

trabalho, no Capítulo 5 são descritos alguns resultados nos quais as influências das dimensões dos volumes de teste ficam evidentes nas imagens de saída, indicando assim a viabilidade de uso da técnica aqui proposta como uma ferramenta de análise de volumes *n*-dimensionais.

Para cumprir a proposta do projeto, foram elencados também objetivos específicos, sendo o primeiro concernente ao ajuste dos valores que compõem os dados de entrada para os processos desenvolvidos. Todos os algoritmos apresentados no Capítulo 4 possuem uma etapa inicial de preparação e ajuste dos dados, de modo que as matrizes possam ser processadas de forma homogênea pelas etapas seguintes de cada um destes algoritmos. Esta etapa é representada pelo processo *Ajuste* no Diagrama de Fluxo de Dados (DFD) que representa a solução proposta.

Apesar de serem permitidos arquivos com diferentes configurações na organização dos dados, e também estarem implementadas algumas rotinas de conversão do conteúdo das matrizes de entrada, todos os arquivos devem apresentar os dados em seu formato bruto, ou seja, com o protótipo desenvolvido não é possível tratar dados provenientes de bases de dados estruturadas, ou mesmo arquivos cujos conteúdos estejam compactados.

Foi especificada também como meta do trabalho a representação dos dados em um espaço de tal modo que os vetores originais pudessem ser apresentados a um algoritmo de visualização. A interpretação da RD, realizada tanto com o algoritmo *Scaling by Mojorizing a Complicated Function* (SMACOF) quanto com a transformação por Coordenadas Estrela, como sendo escalas de cores uni ou bidimensionais, caracteriza a representação dos vetores da matriz de entrada por um conjunto de valores específicos. As representações dos conjuntos utilizados como entrada para o algoritmo de visualização foram apresentados no Capítulo 4, sendo as escalas unidimensionais apresentadas na Figura 24 e as bidimensionais mostradas na Figura 26.

Conforme explicitado no Capítulo 5, o resultado da RD determina quais valores destas escalas são utilizados, pelo algoritmo de visualização, ao calcular a influência de cada um dos vetores que compõem o volume na imagem resultante, atendendo assim o especificado na lista de objetivos. Este resultado corresponde ao processo *RD* no DFD, que descreve a solução apresentada no trabalho.

No Item 3 da lista de objetivos específicos ficou estabelecido também que uma representação visual do conjunto de dados de entrada deveria ser obtida como saída do processo, sendo as imagens, apresentadas nos Capítulos 4 e 5, a demostração de que a visualização de volumes *n*-dimensionais é viável, o que corresponde ao processo *Visualização* no DFD do processo proposto.

Além de sua exequibilidade, pode-se afirmar que as imagens obtidas podem ser utilizadas como uma nova ferramenta de análise de dados, a ser empregada na

investigação de características que podem ser tanto espaciais, quanto de relacionamento entre as diferentes dimensões que compõem o volume de entrada.

Com o intuito de facilitar a análise e interpretação das imagens obtidas a partir da representação do conteúdo de volumes *n*-dimensionais estabeleceu-se como meta a possibilidade de alterar parâmetros do processo de representação visual dos dados. Assim, além da possibilidade de alterar o mapa de cores a ser utilizado no momento da geração das imagens, estão disponíveis também controles para alterar: o brilho, modificando a intensidade das cores apresentadas; a densidade, que ajusta a influência de cada um dos vetores de dados na integral de visualização; a escala da FT, que ajusta a distância entre cada uma das amostras nas escalas de cores; o deslocamento da FT, que altera a posição da primeira amostra nas escalas de cores.

As imagens apresentadas nos Capítulos 4 e 5 têm as configurações de brilho, densidade, escala da FT e ajuste da FT, juntamente com a escolha da escala de cor, ajustadas para melhor representar visualmente o conteúdo de cada dos volumes, além de destacar algumas características obtidas com cada um dos algoritmos implementados para realizar a RD.

Além dos ajustes nos parâmetros de geração das imagens, é possível determinar um recorte nos dados de entrada, conforme apresentado nos exemplos de processamento do volume referente ao Furacão Isabel, que possibilita analisar os vetores de uma região pré determinada do volume de entrada. Complementando a ferramenta de recorte, é possível também, após a RD, escolher uma região do volume processado para gerar a imagem de saída, como o ilustrado na Figura 14. Portanto, é possível afirmar que o objetivo de fornecer ferramentas para alterar os parâmetros do processo de visualização foi atendido.

O último item da lista de objetivos específicos corresponde a executar os algoritmos necessários para a obtenção da representação visual do volume de entrada em um equipamento de baixo custo. Neste sentido o exposto no Capítulo 3, que especifica o uso de um computador pessoal portátil equipado com processador gráfico, atende ao requisito especificado. O uso deste ambiente computacional, entretanto, impôs restrições quanto ao tamanho da matriz que representa o volume de entrada, como o que foi apresentado na Seção 5.2, na qual está explicitado o fato de que a matriz de distâncias necessária para a execução do protótipo utilizando a base de dados do Furacão Isabel possui tamanho $[25.000.000 \times 25.000.000]$, excedendo, em muito, a capacidade de memória dos equipamentos de baixo custo considerados na na realização deste trabalho.

Com exceção da transformação por Coordenadas Estrela, todos os demais algoritmos têm os cálculos matriciais executados na *Graphical Processing Unit* (GPU) do computador, utilizando recursos do ambiente *Compute Unified Device Architec-*

ture (CUDA), descrito na Seção 2.5. Especificamente o processo de geração das imagens é totalmente implementado no processador gráfico, diferentemente dos algoritmos de RD, que possuem parte de sua implementação executada na *Central Processing Unit* (CPU), por apresentarem características lineares de execução.

Estava especificado que uma codificação para representar classes, ou agrupamentos, no volume de entrada deveria estar disponível. Apesar de ser possível distinguir relacionamentos existentes entre as diferentes dimensões da matriz de entrada, conforme exposto no Capítulo 5, não foi implementada uma codificação que garanta a representação de diferentes classes, ou agrupamentos. Contudo, ao alterar a escala de cores em uso, e modificando-se os parâmetros de brilho, densidade e ajustando a escala e o deslocamento dos mapas de cores, é possível distinguir alguns grupos na representação visual produzida.

Durante a execução desta pesquisa, em muitos pontos foram anotadas possibilidades de melhorias tanto com relação ao desempenho dos algoritmos implementados, quanto ao direcionamento de novas investigações e implementações. Baseadas nestas notas, na próxima seção são apresentadas sugestões para a continuação e melhoria deste trabalho.

6.1 SUGESTÕES PARA TRABALHOS FUTUROS

Ao desenvolver o algoritmo de visualização utilizado durante este projeto, percebeu-se que seria possível, além dos mapas de cores uni e bidimensionais apresentados, trabalhar com um volume de cores, de tal forma que para conjuntos de dados com um grande número de dimensões a RD possa ter como destino um espaço tridimensional, melhorando assim as relações de distância entre as projeções, em comparação com o volume original.

Na mesma linha de considerações, as cores representadas nos mapas utilizam o sistema de cores RGB, sendo possível a sua alteração pelo sistema HSI ou HSB. Esta adaptação traria benefícios ao algoritmo de geração das imagens, melhorando a mescla das cores ao calcular a integral de visualização, melhorando, portanto, a representação do volume *n*-dimensional.

Considerando a etapa de RD, conforme explicitado no Capítulo 5, de acordo com as características do volume de dados de entrada, diferentes técnicas devem ser utilizadas, de tal modo a buscar as melhores projeções a serem mapeadas pelo sistema de visualização. Esta característica também é abordada no Capítulo 2, no qual diferentes algoritmos, apresentados na literatura, são comentados. Portanto, em uma ferramenta de análise de dados *n*-dimensionais, será imprescindível a existência de um conjunto maior de alternativas para a realização da RD. A implementação destes deverá seguir a proposta de utilizar o poder de processamento disponível nas GPUs,

disponibilizando assim uma ferramenta de baixo custo aos analistas de dados.

Com relação ao algoritmo SMACOF, é possível realizar a sua implementação de forma tal que os cálculos matriciais sejam realizados com matrizes em bloco (PARK; SHIN; HWANG, 2012), aumentando assim o tamanho do volume de dados que pode ser processado, sem que seja necessário aumentar a quantidade de memória disponível no equipamento. Além do uso das matrizes em bloco, podem ser utilizadas as soluções apresentadas por Pawliczek e Dzwinel (2013) e por Dzwinel e Wcisło (2015), nas quais alterações no método *Multidimensional Scaling* (MDS) possibilitam o processamento com um número reduzido de vetores de entrada, aumentando a velocidade do algoritmo.

O desenvolvimento de uma ferramenta iterativa de análise de dados deve prover, além de algoritmos que preparam os dados de forma a que seja possível ao analista compreender o seu conteúdo (YUAN et al., 2013), como o SMACOF e a transformação por Coordenadas Estrela, ferramentas para manipular tanto o conjunto de entrada quanto o de saída. Agregando algoritmos de agrupamento e classificação de dados às ferramentas já disponíveis no protótipo desenvolvido, pode-se permitir aos analistas, além de selecionar e manipular parcialmente os conjuntos de dados, também identificar e separar estes dados em grupos de acordo com critérios específicos a cada situação.

Além das ferramentas interativas, é importante também disponibilizar algoritmos que realizem a verificação da qualidade dos dados, tanto do volume de entrada, como do resultado obtido na fase de RD.

REFERÊNCIAS

ADHIANTO, L.; BANERJEE, S.; FAGAN, M.; KRENTEL, M.; MARIN, G.; MELLOR-CRUMMEY, J.; TALLENT, N. R. Hpctoolkit: Tools for performance analysis of optimized parallel programs. *Concurrency Computation Practice and Experience*, p. 662–682, 2013. ISSN 15320626.

BAE, S.-H.; QIU, J.; FOX, G. Adaptive interpolation of multidimensional scaling. *Procedia Computer Science*, v. 9, p. 393–402, jan. 2012. ISSN 18770509. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1877050912001639>. Acesso em: 27/02/2014.

BAUMGARTNER, R.; SOMORJAI, R. Graphical display of fmri data : visualizing multi-dimensional space. v. 19, p. 283–286, 2001.

BORG, I.; GROENEN, P. J. F. *Modern Multidimensional Scaling*: Theory and applications. 2. ed. New York, NY: Springer, 2005. ISBN 978-0387-25150-9.

BUZZI, M. F. *Avaliação das correlações de séries temporais de leituras de instrumentos de monitoração geotécnicoestrutural e variáveis ambientais em barragens - Estudo de caso de Itaipu*. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, Abril 2007.

DZWINEL, W.; WCISŁO, R. Very Fast Interactive Visualization of Large Sets of High-dimensional Data. *Procedia - Procedia Computer Science*, Elsevier Masson SAS, v. 51, p. 572–581, 2015. ISSN 1877-0509. Disponível em: <http://dx.doi.org/10.1016/j.procs.2015.05.325>.

ENGEL, D.; HÜTTENBERGER, L.; HAMANN, B. A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization. In: GARTH, C.; MIDDEL, A.; HAGEN, H. (Ed.). *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011*. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012. (OpenAccess Series in Informatics (OASIcs), v. 27), p. 135–149. ISBN 978-3-939897-46-0. ISSN 2190-6807. Disponível em: <http://drops.dagstuhl.de/opus/volltexte/2012/3747>.

ENGEL, K.; HADWIGER, M.; KNISS, J. M.; LEFOHN, A. E.; SALAMA, C. R.; WEISKOPF, D. *Real-time volume graphics*. Wellesley, Massachusetts: A K Peters, Ltd., 2006. 488 p. ISBN 978-1-56881-266-3.

FOLEY, J. D.; DAM, A. van FEINER, S. K.; HUGHES, J. F. *Computer graphics*: Principles and practice. 2. ed. Crawfordsville: Addison-Wesley, 1996. ISBN 0-201-84840-6.

GLASSNER, A. S. *Principles of Digital Image Synthesis*. San Francisco, CA: Morgan Kaufmann Publishers, INC., 1995. 1206 p. ISBN 1-55860-276-3.

GONZALEZ, R. C.; WOODS, R. E. *Processamento Digital de Imagens*. 3. ed. São Paulo: Pearson Prentice Hall, 2010. 624 p. ISBN 978-85-7605-401-6.

GROUP, K. *The open standard for parallel programming of heterogeneous systems*. 2015. Disponível em: <https://www.khronos.org/opencl/>. Acesso em: 28/02/2015.

GUO, H.; XIAO, H.; YUAN, X. Multi-dimensional transfer function design based on flexible dimension projection embedded in parallel coordinates. In: IEEE. *IEEE Pacific Visualization Symposium*. 2011. p. 19–26. Disponível em: <http://vis.pku.edu.cn/research/publication/pacificvis11-hdtf-small.pdf>.

ISENBERG, T.; ISENBERG, P.; CHEN, J.; SEDLMAIR, M.; MOLLER, T. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, v. 19, n. 12, p. 2818–2827, 2013. ISSN 10772626.

ITAIPU BINACIONAL. *Usina Hidrelétrica de Itaipu: Aspectos de engenharia*. Foz do Iguaçu, Pr: Itaipu Binacional, 2009.

KANDOGAN, E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In: CITESEER. *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics*. Salt Lake City, Utah, 2000.

KREUSELER, M. Visualization of geographically related multidimensional data in virtual 3d scenes. *Computers & Geosciences*, v. 26, n. 1, p. 101–108, 2000. ISSN 0098-3004. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098300499000369>. Acesso em: 09/07/2014.

KURZWEIL, R. *The age of spiritual Machines*. 1. ed. New York: Viking Penguin, 1999. ISBN 0-670-88217-8.

LAWRENCE, J.; ARIETTA, S.; KAZHDAN, M.; LEPAGE, D.; O'HAGAN, C. A user-assisted approach to visualizing multidimensional images. *IEEE transactions on visualization and computer graphics*, v. 17, n. 10, p. 1487–98, out. 2011. ISSN 1941-0506. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21817169>. Acesso em: 09/07/2014.

MARTINS, R. M.; COIMBRA, D. B.; MINGHIM, R.; TELEA, a. C. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers and Graphics (Pergamon)*, Elsevier, v. 41, n. 1, p. 26–42, 2014. ISSN 00978493. Disponível em: <http://dx.doi.org/10.1016/j.cag.2014.01.006>. Acesso em: 21/09/2015.

MATOS, S. F. *Avaliação de instrumentos para auscultação de barragem de concreto. Estudo de caso: Deformímetros e tensômetros para concreto na barragem de Itaipu*. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, 2002.

MATRAKAS, M. D.; SCHEER, S. Three-Dimensional Representation of a Multidimensional Data Set. *Applied Mathematical Sciences*, HIKARI Ltd, v. 10, n. 20, p. 959–971, 2016.

MATRAKAS, M. D.; SCHEER, S. Visualization of Multidimensional Data Volumes Using Dimensional Reduction Techniques. *International Journal of Engineering and Applied Sciences*, EAAS & ARF, v. 9, n. 01, p. 12 – 20, 2016.

MOKBEL, B.; LUEKS, W.; GISBRECHT, A.; HAMMER, B. Visualizing the quality of dimensionality reduction. *Neurocomputing*, Elsevier, v. 112, p. 109–123, 2013. ISSN 09252312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2012.11.046>. Acesso em: 21/09/2015.

NAJIM, S. a.; LIM, I. S. Trustworthy dimension reduction for visualization different data sets. *Information Sciences*, Elsevier Inc., v. 278, p. 206–220, 2014. ISSN 00200255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2014.03.048>. Acesso em: 21/09/2015.

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR). *Hurricane Isabel WRF Model Data*. 2009. Disponível em: <http://www.vets.ucar.edu/vg/isabeldata/readme.html>.

NELSON, B.; KIRBY, R. M.; HAIMES, R. GPU-based volume visualization from high-order finite element fields. *IEEE transactions on visualization and computer graphics*, v. 20, n. 1, p. 70–83, jan. 2014. ISSN 1941-0506. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24201327>. Acesso em: 09/07/2014.

NVIDIA. *NVidias's next generation CUDA compute architecture: FERMI*. 2009. Disponível em: <http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf>. Acesso em: 10/12/2014.

NVIDIA. *CUDA C Programming Guide*. 2014. Disponível em: <http://docs.nvidia.com/cuda/cuda-c-programming-guide>. Acesso em: 05/08/2014.

NVIDIA. *CUDA Toolkit Documentation*. 2014. Disponível em: <http://docs.nvidia.com/cuda>. Acesso em: 05/08/2014.

PAO, Y.-H.; MENG, Z. Visualization and the understanding of multidimensional data. *Engineering Applications of Artificial Intelligence*, v. 11, n. 5, p. 659–667, 1998. ISSN 0952-1976. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0952197698000311>. Acesso em: 09/07/2014.

PARK, S.; SHIN, S. yong; HWANG, K. baek. Cfmds : Cuda-based fast multidimensional scaling for genome-scale data. *BMC Bioinformatics*, BioMed Central Ltd, v. 13, n. Suppl 17, p. S23, 2012. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/13/S17/S23>. Acesso em: 09/07/2014.

PATIAS, J. *Zoneamento geotécnico com base em Krigagem ordinária e equações multiquádricas: Barragem de Itaipu*. Tese (Doutorado) — Universidade de São Paulo - USP, São Carlos - SP, Outubro 2010.

PATTERSON, D. A.; HENNESSY, J. L. *Organização e projeto de computadores*: a interface hardware/software. 3. ed. [S.l.]: Elsevier, 2005. ISBN 85-352-1521-2.

PAWLICZEK, P.; DZWINEL, W. Interactive data mining by using multidimensional scaling. *Procedia Computer Science*, Elsevier B.V., v. 18, p. 40–49, 2013. ISSN 1877-0509. Disponível em: <http://dx.doi.org/10.1016/j.procs.2013.05.167>.

PERES, R. T.; ARANHA, C.; PEDREIRA, C. E. Optimized bi-dimensional data projection for clustering visualization. *Information Sciences*, v. 232, p. 104–115, maio 2013. ISSN 00200255. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0020025513000339>. Acesso em: 27/02/2014.

SANTOS, S. dos; BRODLIE, K. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, v. 28, n. 3, p. 311–325, jun. 2004. ISSN 00978493. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0097849304000251>. Acesso em: 27/02/2014.

SHIEH, S.-L.; LIAO, I.-E. A new approach for data clustering and visualization using self-organizing maps. *Expert Systems with Applications*, Elsevier Ltd, v. 39, n. 15, p. 11924–11933, nov. 2012. ISSN 09574174. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417412004447>. Acesso em: 09/07/2014.

TANENBAUM, A. S. *Organização Estruturada de Computadores*. 5. ed. São Paulo, SP: Pearson Prentice Hall, 2007. ISBN 978-85-7605-067-4.

TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, v. 290, n. 5500, p. 2319–23, 2000. ISSN 0036-8075. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11125149>. Acesso em: 11/10/2015.

TSAI, F. S. A visualization metric for dimensionality reduction. *Expert Systems with Applications*, Elsevier Ltd, v. 39, n. 2, p. 1747–1752, fev. 2012. ISSN 09574174. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417411011924>. Acesso em: 27/02/2014.

TUKEY, J. *Exploratory data analysis*. [S.l.]: Addison-Wesley, 1977.

van der MAATEN, L.; POSTMA, E.; van den HERIK, J. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, v. 10, p. 1–41, 2009. ISSN 0169328X.

WARD, M. O.; GRINSTEIN, G.; KEIM, D. *Interactive Data Visualization*: Foundations, techniques and applications. Boca Raton, Fl: CRC Press, 2015. 558 p. ISBN 978-1-4822-5737-3.

WIKIPEDIA. *Moore's law*. 2015. Disponível em: <https://en.wikipedia.org/wiki/Moore\%27s_law>.

WRIGHT, H. *Introduction to Scientific Visualization*. 1. ed. Hull - UK: Springer, 2007. 147 p. ISBN 978-1-84628-494-6.

YUAN, X.; REN, D.; WANG, Z.; GUO, C. Dimension projection-matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. v. 19, p. 2625–2633, 2013. Disponível em: <http://vis.pku.edu.cn/research/publication/infovis13_dpmatrixtree.pdf>. Acesso em: 02/09/2014.